USENIX Association

Proceedings of the Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)

August 9–10, 2021

ISBN 978-1-939133-25-0

Symposium Organizers

General Chair
Sonia Chiasson, *Carleton University*

Technical Papers Co-Chairs
Joe Calandrino, *Federal Trade Commission*
Manya Sleeper, *Google*

Technical Papers Committee
Yasemin Acar, *Max Planck Institute for Security and Privacy*
Florian Alt, *Bundeswehr University Munich*
Olabode Anise, *Google*
Adam J. Aviv, *The George Washington University*
Rebecca Balebako, *Google*
Lujo Bauer, *Carnegie Mellon University*
Jasmine Bowers, *MITRE*
Cristian Bravo-Lillo, *Ciberseguridad Humana*
Lynne Coventry, *Northumbria University*
Lorrie Cranor, *Carnegie Mellon University*
Carrie Gates, *Bank of America*
Maximilian Golla, *Max Planck Institute for Security and Privacy*
Julie Haney, *National Institute of Standards and Technology (NIST)*
Marian Harbach, *Google*
Jun Ho Huh, *Samsung Research*
Iulia Ion, *Google*
Hassan Khan, *University of Guelph*
Bart Knijnenburg, *Clemson University*
Katharina Krombholz, *CISPA Helmholtz Center for Information Security*
Janne Lindqvist, *Aalto University*
Shri Mare, *Western Washington University*
Abigail Marsh, *Macalester College*
Susan E. McGregor, *The Data Science Institute at Columbia University*
Mainack Mondal, *Indian Institute of Technology*
Xinru Page, *Brigham Young University*
Irwin Reyes, *Two Six Labs*
Scott Ruoti, *The University of Tennessee*
Florian Schaub, *University of Michigan*
Kent Seamons, *Brigham Young University*
Elizabeth Stobert, *Carleton University*
Jose M. Such, *King's College London*
Eran Toch, *Tel Aviv University*
Blase Ur, *University of Chicago*
Kami Vaniea, *University of Edinburgh*
Emanuel von Zezschwitz, *Google*
Yang Wang, *University of Illinois at Urbana–Champaign*
Rick Wash, *Michigan State University*
Josephine Wolff, *Tufts University*
Leah Zhang-Kennedy, *University of Waterloo*
Mary Ellen Zurko, *MIT Lincoln Laboratory*

Lightning Talks and Demos Co-Chairs
Sanchari Das, *University of Denver*
Blase Ur, *University of Chicago*

Lightning Talks and Demos Junior Co-Chair
Amel Bourdoucen, *Aalto University*

Karat Award Chair
Nalin Asanka Gamagedara Arachchilage, *La Trobe University*

Posters Co-Chairs
Camille Cobb, *Carnegie Mellon University*
Maximilian Golla, *Max Planck Institute for Security and Privacy*

Posters Junior Co-Chair
Yixin Zou, *University of Michigan*

Tutorials and Workshops Co-Chairs
Katharina Krombholz, *CISPA Helmholtz Center for Information Security*
Leah Zhang-Kennedy, *University of Waterloo*

Tutorials and Workshops Junior Co-Chair
Imani Sherman, *University of Florida*

Mentoring Co-Chairs
Adam J. Aviv, *The George Washington University*
Mohamed Khamis, *University of Glasgow*

Mentoring Junior Co-Chairs
Ruba Abu-Salma, *International Computer Science Institute (ICSI) and University of California, Berkeley*
Borke Obada-Obieh, *University of British Columbia*

Publicity Co-Chairs
Charles Weir, *Lancaster University*
Yaxing Yao, *Carnegie Mellon University*

Email List Chair
Lorrie Cranor, *Carnegie Mellon University*

Accessibility Chair
Liz Markel, *USENIX Association*

USENIX Liaison
Casey Henderson, *USENIX Association*

External Reviewers

David Balash
Amel Bourdoucen
Pam Briggs
Sonia Chiasson
Geumhwan Cho
James Conners
Florian Farke
Matthias Fassl
Reza Ghaiumy Anaraky
Franziska Herbert
Maritza Johnson
Smirity Kaushik
Patrick Kelley
Marvin Kowalewski
Xi Lu
Rongjun Ma
Sana Maqsood
Philipp Markert
Michelle Mazurek
Jaron Mink
Moses Namara
Katharina Pfeffer
Emilee Rader
Rob Reeder
Tanusree Sharma
WooChul Shim
Garrett Smith
Noel Warford
Yaxing Yao
Daniel Zappala
Zhuohao Zhang
Zhixuan Zhou

Message from the SOUPS 2021 Program Co-Chairs

Welcome to SOUPS 2021!

With the conference in its 17th year, our SOUPS community has collectively ensured an excellent and exciting conference program despite the challenges and obstacles caused by the global pandemic. With a record 36 papers accepted out of 136 submissions (26% acceptance rate), the technical program covers a wide range of topics within usable privacy and security. The conference also includes workshops, posters, lightning talks, mentorship activities, and a keynote.

In 2016, SOUPS became an independent conference body. For the last five years, we have partnered with USENIX for hosting and administrative support, a move that has enabled continued growth for the conference. We thank all the members of the USENIX staff for their work in organizing SOUPS and supporting our community. We particularly appreciate their support and flexibility this year, including managing the virtual event. Their team has been fantastic at making the process seamless. In 2018, we co-located with the USENIX Security Symposium for the first time, and we have continued that co-location virtually for 2021. Co-locating the two conferences allows for interactions and shared ideas between SOUPS and USENIX Security attendees. We have found this beneficial for both conferences and look forward to the opportunity again this year. While we miss in-person interactions, a virtual conference format has also facilitated participation for many, and we believe that this can lead to richer discussions and wider perspectives within the community. We hope that you will find SOUPS 2021 engaging and meaningful.

SOUPS relies on a range of volunteers for all of its activities. Steering Committee members provide oversight and guidance and are elected for three-year terms. Organizing Committee members help determine the conference content for a particular year, often serving two-year terms to facilitate the transition of knowledge. This year, we introduced corresponding junior co-chairs for each of our chairing roles within the Organizing Committee, to offer senior students or early career researchers a voice in the organization of SOUPS and to prepare them to take on leadership roles in upcoming years, whether with SOUPS or other conferences. Technical Papers Committee members are chosen by the Technical Papers Co-Chairs each year. This year, the on-going pandemic continued to create obstacles for many of us, including members of the Technical Papers Committee. As a result, other members of the committee and external reviewers stepped in to assist with reviews. SOUPS is a product of the hard work by many people, starting with researchers who decide to submit their work to SOUPS, and including all of the SOUPS Organizers, the SOUPS Steering Committee, the technical paper reviewers, the workshop organizers, the poster jury, and the USENIX staff. We are grateful and thank each and every one of you for your contributions to SOUPS 2021.

Sonia is serving as General Chair of SOUPS and Chair of the Steering Committee for 2021 and 2022. A Vice Chair will be appointed for 2022, who will then take on the role of General Chair for the following two years. If you are interested in helping with SOUPS 2022 in any way, please contact Sonia.

SOUPS would not be possible without the generous support of our sponsors – thank you. Please visit our website to view the recipients of the SOUPS 2021 awards. Congratulations to all recipients for their outstanding work.

Sonia Chiasson, *Carleton University*
General Chair

Joe Calandrino, *Federal Trade Commission*
Technical Papers Co-Chair

Manya Sleeper, *Google*
Technical Papers Co-Chair

Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)

August 9–10, 2021

Monday, August 9

Authentication

User Attitudes and (Mis)understandings

Perspectives and Policy

Inclusive Privacy and Security

Tuesday, August 10

Phishing and Account Compromise

Security in Practice

Ubiquitous Computing

Developers

Work (and Learn) from Home

Towards Usable and Secure Location-based Smartphone Authentication

Geumhwan Cho
Sungkyunkwan University

Sungsu Kwag
Samsung Research

Jun Ho Huh
Samsung Research

Bedeuro Kim
Sungkyunkwan University

Choong-Hoon Lee
Samsung Research

Hyoungshick Kim
Sungkyunkwan University

Abstract

The concept of using location information to unlock smartphones is widely available on Android phones. To date, however, not much research has been conducted on investigating security and usability requirements for designing such location-based authentication services. To bridge this gap, we interviewed 18 participants, studying users' perceptions and identifying key design requirements such as the need to support fine-grained indoor location registration and location (unlock coverage) size adjustment. We then conducted a field study with 29 participants and a fully-functioning application to study real-world usage behaviors. On average, the participants were able to reduce about 36% of manual unlock attempts by using our application for three weeks. 28 participants enduringly used registered locations to unlock their phones despite being able to delete them during the study and unlock manually instead. Worryingly, however, 23 participants registered at least one insecure location – defined as a location where an unwanted adversary can physically access their phones – as a trusted location mainly due to convenience or low (perceived) likelihood of phones being attacked. 52 out of 65 total registered locations were classified as insecure by the definition above. Interestingly, regardless of whether locations were considered secure or insecure, the participants preferred to select large phone unlock coverage areas.

1 Introduction

Users' location information can be used as an additional factor to improve authentication security or usability [26]. For instance, users' daily location traits can be trained and used to detect anomalous use of smartphones. Banks detect financial frauds in a similar way [19,22]. In a risk-based authentication scheme [14], users may be allowed to use certain services without using explicit authentication if they are logging in from a secure location.

In 2014, Google launched an automatic phone unlock scheme for Android called "Smart Lock" [1]. One of its features allows users to freely select "trusted places" to automatically unlock phones, and keep them unlocked while users are using their phones within secure locations that are supposed to be safe from unauthorized access. Smart Lock's trusted places feature relies mainly on GPS to detect users' trusted locations. As a result, Google estimates that phones may remain unlocked within a radius of up to about 80 meters (from the registered spot) [1] – specifying a fine-grained indoor location area is almost infeasible. Users cannot customize trusted location sizes – there is no option to reduce or increase location sizes. Such limitations may raise security and usability concerns for users, and discourage them from adopting this scheme [20]. In this paper, we focus on this specific notion of unlocking smartphones based on location information – identifying key design requirements, gauging real-world usability benefits related to reducing manual (explicit) phone unlock burden, and analyzing potential security issues that could arise from freely allowing users to select trusted places.

The use of location information to unlock phones implies that we are treating this information – i.e., the physical security offered by trusted locations – to provide a comparable security level to those provided by existing screen unlock schemes. This assumption could potentially put smartphone users at severe risks of phone breaches. For instance, a user with low-security awareness might register public locations such as cafes or sports facilities for convenience. An adversary who manages to steal that phone would be able to easily unlock it by just going near those locations.

We first conducted an interview study with 18 participants to understand users' perceptions and expectations on

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

location-based smartphone authentication. We then developed a location-based screen unlock application for Android based on the design requirements identified from the first study, and conducted a real-world field study with 29 participants. The reason we developed our own application was to reuse the Smart Lock concepts (since this is the only known real-world application) while also providing support for indoor location detection based on WiFi RSSI information. After obtaining informed consent, we asked the participants to install our application on their own phones and use it for three weeks; we logged the participants' real-world usage behaviors. The key observations made from analyzing this data and paper contributions are summarized below:

- We identified security and usability requirements for developing location-based authentication systems through the first study: these requirements include the need to support fine-grained indoor location registration, and allow users to select and adjust location coverage sizes. The field study results confirmed that people indeed register indoor locations (e.g., homes and offices), and choose different location sizes.
- Using a fully functional application (implemented based on the requirements), we conducted a three-week field study to collect real-world usage data. Our findings indicate that the location-based automatic unlock feature would be immensely beneficial – the participants, on average, were able to reduce about 36% of their explicit unlock attempts.
- Even though the participants were free to delete all locations during the study (and go back to manual unlocks), 28 out of 29 participants continued using at least one location throughout the study. During the post-study interview, 22 participants said that they would continue using our application due to the automatic phone unlock convenience. These observations highlight the usability benefits.
- Worryingly, we identified two critical security issues: (1) many users have a tendency to register insecure locations (defined as locations that are vulnerable to unauthorized phone access) – 52 out of 65 registered locations were considered potentially insecure; and (2) regardless of whether locations are considered secure or insecure, the participants preferred to select large location coverage sizes.

2 Related work

The concept of using location information for authentication was first introduced by Denning and MacDoran [7]. The key idea is to use a user's physical location information as an additional factor to verify the validity of log in requests. Numerous existing studies [14, 19, 22] have applied this idea to improve authentication security by verifying users' known locations. For example, banks may compare users' phone locations and the payment terminal locations to detect frauds [19, 22]. Daniel et al. [14] proposed a location-based risk assessment framework to facilitate automatic adjustment of required authentication factors (steps) based on risk levels.

Several studies [4, 10, 15, 18] have demonstrated that users' physical location traits can be unique and be used to identify users. Fridman et al. [10] demonstrated that device location information could be used to identify users – using GPS coordinates as the main classification features, they were able to identify users with an FAR and an FRR below 0.1 and 0.05, respectively. Agadakos et al. [4] proposed a location-based authentication method that analyzes proximity information between users' phones and paired IoT devices. Two recent studies [5, 16] proposed phone theft detection techniques based on the use of acoustic signals to measure physical distance between users and phones. Li et al. [16] used frequency-modulated carrier waves to measure physical distances. Chen et al. [5] used information about users' motions to improve the accuracy of measuring distances.

An accurate algorithm for determining users' locations is essential in implementing location-based authentication. Several studies (e.g., [6]) discussed the use of wireless (e.g., WiFi) signals to identify device locations. Hilsenbeck et al. [13] presented a fusion approach using sensors: they were able to track a user with 1.52m accuracy 50% of the time, and 4.53m accuracy 90% of the time. Shu et al. [25] presented another fusion approach using magnetic and WiFi signals to achieve 3.5m accuracy 90% of the time. Abbas et al. [3] proposed a deep learning-based indoor localization technique to achieve 2.38m accuracy 50% of the time in a university building. Such techniques focus on developing classification models for accurate location detection. In this paper, we implemented a fully working indoor location-based authentication solution that does not require special hardware or a fingerprinted wireless signal map.

Mehrabi et al. [20] investigated how users perceive Smart Lock and its trusted places feature [1]. Their surveys, however, primarily focus on understanding why people are willing to use or not use the trusted places feature. We dived deep into understanding users' specific functional needs and expected behaviors to derive application design requirements. Moreover, through the field study, we investigated the real-world effectiveness of location-based automatic unlock schemes and identified security issues that need to be mitigated.

3 Requirement Study

3.1 Methodology

As the first step, we conducted a semi-structured interview study to understand users' perceptions and expectations with respect to the use of trusted physical locations to unlock their phones implicitly. We recruited 18 participants who are aged 18 years or older by posting advertisements on online notice boards at a university as well as selectively recruiting people from local communities based on their age and work expe-

riences to ensure that overall demographic proportions are similar to those presented in [2]. Two moderators together ensured that all of the interview questions were asked and consistently understood by the participants. Each study session took about 20 minutes on average to complete, and participants were compensated for their time with a USD 10 gift card. All interviews were recorded and transcribed.

As for all open-ended questions, we applied structural coding techniques [17] [24] to identify responses to each interview question on transcripts, and 24 topic codes were identified through thematic coding. One researcher was the primary coder, responsible for creating and updating the codebook. The other two researchers independently coded interview transcripts, revised the codebook, and resolved disagreements. After resolving coding disagreements, we achieved inter-coder agreement of 89% Cohen's Kappa [9].

The participants were informed that participation is voluntary and confidential, and they have the right to terminate the study without penalty. We asked for their permission to audio-record entire interview sessions. The ethical perspective of the requirement study was validated through an institutional review board (IRB) at a university.

Before asking questions, the interviewers explained the basic concept of registering trusted locations, and using those registered locations to automatically unlock phones. We borrowed the exact instruction phrase from Smart Lock, which says "*Add location where device should be unlocked.*"

We then asked participants three simple questions about how this authentication service would work in practice (e.g., "*What happens to your phone when you physically move to a place that you already registered as a trusted location?*") to ensure that all participants had an adequate level of understanding of this concept before the interview. For those who answered any of the three questions wrong, we spent more time explaining this concept until they were comfortable with it.

The interview questions are as follows: The first question we asked was "*Provide a list of places that you would register as a trusted location and explain why.*" We then asked the participants to "*Select a size (that defines the area in which their phones would remain unlocked) for each of your trusted locations, and explain why.*" The participants were also asked to explain what would be a tolerable setup time (i.e., time taken to register one location), battery consumption level, and location detection accuracy.

Before conducting the interview, we conducted a pilot study with 3 participants and used their feedback to revise the study structure, interview questions, and guidelines.

3.2 Results

3.2.1 Demographics

We interviewed a total of 18 participants. 10 out of 18 were females, and the average age was 39.1 (σ = 11.6). 9 participants had a university degree, and 6 participants had a master (or doctoral) degree. 13 participants said they unlock their phones many times an hour. 15 participants said they store sensitive or confidential information on their phones. 9 different occupations were reported with "personal care and service occupations," "student," "education, training, and library occupations," and "management occupations" being the top ones. Only one participant used Smart Lock and registered home as a trusted place. To demonstrate the representativeness of the samples, we compared the age, gender, and education distributions against the US smartphone population reported in [2]. We used Fisher's exact tests to show that there are no significant differences in age ($p = 0.26$), gender ($p = 0.64$), and education ($p = 0.18$) distributions. The details of demographics are summarized in Appendix A. We performed data collection and analyses concurrently until we reached theoretical saturation. Figure 1 shows the code saturation results. There are no new codes between 17th and 18th participants. The number of codes reported in the requirement study is 23 in total.

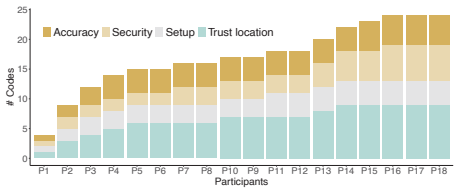


Figure 1: Code saturation results. "Accuracy" indicates unlock accuracy expectations; "Security" indicates security expectations; "Setup" indicates setup time; and "Trust location" indicates trust location considerations.

3.2.2 Trusted Location Considerations

The first question we asked was *"What physical locations or places would you register as trusted locations and allow your phone to be unlocked automatically? Explain why."* Table 1 shows different types of physical locations that the participants consider as trusted, and provides the number of times each location was mentioned. 6 out of 18 participants mentioned three different locations, 9 participants mentioned two different locations, and 3 participants mentioned one location. Unsurprisingly, "home" was the most frequently mentioned trusted location, followed by "office," and "my room."

As for the reasons for selecting trusted locations, we identified 7 different codes. Note that some participants provided multiple reasons. The most frequently cited reasons were `private space` and `frequently visited place`, each of which was mentioned by 7 participants. P1 mentioned "my room" and the privacy it offers:

Table 1: Types of trusted locations, and counts for each location type. Rows "One," "Two," and "Three" refer to the number of locations that each participant mentioned as trusted locations; for instance, row "Three (6)" indicates there were six participants who each mentioned three different locations.

# Locations (# Participants)	One (3)	Two (9)	Three (6)	**Total (18)**
Home	3	8	5	**16**
Office	0	7	3	**10**
My room	0	1	4	**5**
Office desk	0	1	1	**2**
Lecture room	0	0	2	**2**
Church	0	1	0	**1**
Bathroom	0	0	1	**1**
Cafe	0	0	1	**1**
Gym	0	0	1	**1**
Total	**3**	**18**	**18**	**39**

> *"My room... It's completely my own space. Even if I'm at home, there are things that I do not want to share with my family.."* (P1)

Another frequently cited reason was `spend a lot of time`, which was mentioned by 3 participants. P12 mentioned "home," because he spends most of the time at "home", and would like the phone to remain unlocked while he is at "home." P16 specifically mentioned that she registered "church" because she believes it is a `trustworthy place`.

3.2.3 Trusted Location Sizes

The participants were asked *"If you were able to specify a radius of a circle to indicate the size of a trusted location you mentioned earlier, what would be a radius size that you prefer? Answer in meters."* This question was designed to gauge users' preferences with respect to specifying trusted location coverage sizes.

Table 2: Numbers of preferred trusted location coverage sizes in meters for each location type.

Location	1–3m	4–6m	7–9m	10–12m	13–15m
Home	2	2	4	8	0
Office	1	6	2	1	0
My room	3	2	0	0	0
Office desk	2	0	0	0	0
Lecture room	0	0	0	1	1
Church	0	0	0	0	1
Bathroom	1	0	0	0	0
Cafe	0	0	0	0	1
Gym	0	0	0	0	1
Total	**9**	**10**	**6**	**10**	**4**

Table 2 shows the coverage sizes that users preferred for each location type. Smaller sizes, less than 6 meters, were mostly preferred for individual rooms and offices. P6 said he would like the phone to remain unlocked only when he is working at the desk. Larger sizes, larger than 7 meters, were preferred for homes. P3 mentioned that she trusts the entire space of her home and does not mind the phone being unlocked in her home. As for all the public (freely accessible) locations that were mentioned (lecture room, church, cafe, and gym), the participants preferred larger sizes – this observation raises potential security concerns. These observations indicate that location-based authentication services should allow users to select different location sizes.

3.2.4 Setup Time

To gauge what range of setup times users are willing to tolerate when registering trusted locations, we asked *"What do you consider to be an adequate time taken to register one trusted location (answer in seconds or minutes)?"* The average setup time the participants were willing to tolerate was 3.2 minutes ($\sigma = 2.5$). 7 participants emphasized that setup times need to be short. One response was:

> *"About one minute. If the setup time is too long I will not use it."* (P6)

Two participants mentioned that the setup times should be similar to that of setting up other unlock options like patterns or PINs. Here is a quote from P14:

> *"I don't want to use up more time than what I would normally spend setting up a pattern."* (P14)

3.2.5 Unlock Accuracy Expectations

To understand users' location detection accuracy expectations, we asked *"A location-based authentication error occurs when it fails to unlock your phone when you physically move to a registered trusted location. How many failures out of 10 attempts are you willing to tolerate before stopping the use of a location-based authentication service?"* 2 out of 18 participants mentioned they would not tolerate any unlock failure. 6 participants said they would tolerate just one failure. P9 mentioned:

> *"..it's impossible to have zero failure.. one [out of ten] failure would not be that inconvenient.."* (P9)

4 participants mentioned that they would tolerate two failures. 2 participants were willing to tolerate three failures. 4 participants said they would tolerate five or six failures. P14 was willing to tolerate 5 failures:

> *"..five.. current unlock methods also frequently fail anyway.."* (P14)

Overall, we observed a wide range of failure tolerance levels among the participants, ranging between 0 to 6 (out of 10 unlock attempts) failures. However, the majority of the participants expected one or two failures.

3.2.6 Security Expectations

Similarly, to understand the participants' security expectations, we asked *"A location-based authentication security failure occurs when it fails to lock your phone after physically walking away from registered trusted locations. How many security failures out of 10 attempts are you willing to tolerate before stopping the use of a location-based authentication service?"* The participants were more strict with security: 6 out of 18 participants mentioned that they would not tolerate any security failure. P17 mentioned:

> *"Because this technology is about automatically unlocking my phone, it needs to guarantee high [location detection] accuracy.."* (P17)

9 participants said they would tolerate one or two security failures. However, there were more participants (compared to those who were unwilling to tolerate any unlock failure) who expected no security failure.

3.2.7 Battery Use

To understand what level of battery use the participants are willing to tolerate, we asked *"How much battery use are you willing to tolerate before stopping the use of a location-based authentication service?"* The distribution of responses indicates that tolerable battery usage percentage per day mainly ranged from 5 to 15%. (see Appendix B).

3.3 Requirements

Based on the above observations, we summarize key design requirements that must be considered upon designing a usable and secure location-based authentication service:

1. **Indoor locations.** Many participants expressed their preferences to register indoor locations such as rooms and offices as trusted locations – the first requirement is that a service should allow users to register indoor locations as trusted locations.

2. **Multiple locations.** Except for one participant, everyone expressed the preference to register two or more trusted locations. The second requirement is that a service should allow users to register more than one trusted location.

3. **Adjustable location sizes.** The participants expressed different location coverage preferences. The third requirement is that a service should allow users to choose different location coverage sizes and adjust them individually.

4. **Setup time.** Based on responses about tolerable setup times, the fourth requirement is that users should be able to register a single location within 3.2 minutes.

5. **False rejection rates and false acceptance rates.** The majority of the participants said they were willing to tolerate one or two security/lock failures for every ten lock attempts (phones remaining unlocked when users move away from trusted locations). This error rate is referred to as false acceptance rates (FARs). Similarly, most were willing to tolerate one or two usability/unlock failures for every 10 unlock attempts (phones remaining locked when users try to use them inside trusted locations). The error rate is referred to as false rejection rates (FRRs). Such tolerable lock or unlock failure levels need to be satisfied at the minimum.

6. **Battery use.** The participants were willing to tolerate between 5 to 15% use of battery during daytime for running a location-based authentication service.

3.4 Limitations

In the requirement study, a small number of participants may not be sufficient to enumerate all possible codes to understand the requirements for location-based authentication. To address this issue, we tested whether code saturation was reached with two separate coders.

Moreover, the participants could have possibly misunderstood some of the questions/terms because all participants except one participant who has used Smart Lock did not use any location-based authentication scheme before the study. For example, the term of trusted location can be differently interpreted by each participant. To keep the chances of such misunderstanding low and ensure consistency, we had two researchers interviewing together in the requirement study and conducted a pilot study before the requirement study to resolve the ambiguity and misconceptions surrounding the terms and questions.

Since our studies were designed to use self-reported data, our results inherently depend on the participants' honesty and knowledge. We mitigated this limitation by conducting the field study with a fully working Android application that supports location-based authentication.

4 Field Study Application Design

As the next step, we implemented a fully functional location-based authentication application that follows the phone lock/unlock paradigms introduced through Smart Lock yet also contains new features that we identified as important through the first study. We used this application to conduct a field study and analyze users' real-world usage behaviors.

4.1 Design Overview

We named our location-based smartphone authentication application "Loclock". Because the GPS technology alone is not sufficient to support the first "indoor locations" requirement,

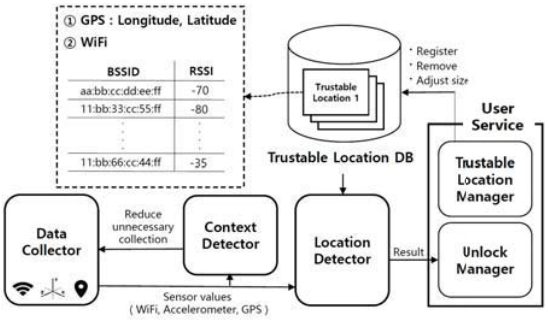


Figure 2: Overview of Loclock.

we also used WiFi information – more specifically, signal strengths of nearby access points – to create fingerprints for indoor locations. To satisfy the "adjustable location sizes" requirement, we designed Loclock to support three different location coverage sizes. Since we cannot guarantee meter-level location detection accuracy, we provide three coverage options that users can choose from: 0 to 5 meters, 5 to 10 meters, and 10 or more meters. Figure 2 shows the architectural overview of Loclock.

Data Collector. To satisfy the "battery use" requirement, we tried to minimize the number of sensors used for collecting data. We collect accelerometer sensor data, GPS data, and the WiFi "received signal strength indication" (RSSI) values from nearby access points. GPS data are used for large area (usually outdoor) detection, and WiFi RSSI values are used for more fine-grained indoor area detection. Accelerometer data are used for context detection.

Context Detector. The accelerometer data are used to detect when a phone is sitting idle on a specific place (e.g., desk). We use this contextual information to determine when to stop or start collecting WiFi RSSI values because continuous and frequent WiFi RSSI collection would use up too much battery. For instance, when a user leaves her phone on her desk, there is no need to collect WiFi RSSI values frequently while the phone is sitting idle on the desk. We measured battery consumption levels in a lab setting for intensive and less intensive battery use scenarios. The intensive battery use scenario collected all sensor data and WiFi signals, but the less intensive scenario collected sensor data only. Our evaluation results showed that the first scenario consumed about 9 percent per hour, while the less intensive scenario consumed about 3 percent per hour.

Location Detector. This component detects whether a phone is inside a registered location coverage area. As the first step, GPS information is used to determine whether a registered location is inside a large coverage area. To avoid unnecessary battery drain, in the case when the phone is inside the large coverage area, it collects WiFi RSSI values from the nearby access points of the current location and compares them against pre-stored (upon trusted location registration) RSSI values. WiFi RSSI values could be sensitive and differently measured under various environmental conditions. When a user stores the RSSI values for a trusted location during the trusted location registration process, we found that one minute is reasonable to collect a sufficient number of RSSI values while satisfying the "setup time" requirement. Loclock uses the average value for each access point to avoid the bias by some outlier RSSI values. The lower Euclidean distance between current WiFi RSSI values and pre-stored RSSI values (upon trusted location registration), the closer the current location is to a registered trusted location. We set a distance threshold to determine whether the phone is inside a trusted location coverage area: if a distance value is lower than the threshold – this indicates that a given location is a trusted location – the phone will be unlocked. We empirically determined the optimal threshold.

User Service. This component allows users to configure PIN, pattern, or password as a screen unlock scheme. Users must set up at least one scheme before using Loclock. Such schemes are used to unlock phones when users are not inside trusted location coverage areas, or when Loclock fails to unlock phones inside trusted locations. This component also provides the user interface for users to register, modify, or delete trusted locations.

4.2 Lock/Unlock Failure Rate Evaluation

To demonstrate that Loclock can achieve tolerable failure rates as described in the "FRR and FAR" requirement, we collected WiFi RSSI datasets from three different locations (two office buildings and one university laboratory) using Loclock and evaluated the lock and unlock failure rates with varying threshold values. We provide a summary of the evaluation results in Appendix C.

For each of the two coverage sizes, 5 and 10 meters, we measured three sets for FRR and FAR, fixing FRRs to 10, 20, and 30% – this would give us three specific RSSI threshold values that guarantee those three FRR rates – and measuring three FARs based on the three threshold values. At both FRR 10 and 20% threshold values, the FARs were contained around 20%. The half total error rates (HTER), computed by averaging FARs and FRRs, are all below 20% when FRRs are fixed at 10 and 20%. Referring back to the "FRR and FAR" requirement (willing to tolerate one or two out of 10 failures), these FRR/FAR results indicate the field study participants would likely experience tolerable error rates.

5 Field Study

We designed the second field study based on the observations from the requirement study. The majority of the participants hypothetically selected at least two different trusted locations, mentioning various indoor location types ranging from homes to public places like cafes or gyms. Through the field study,

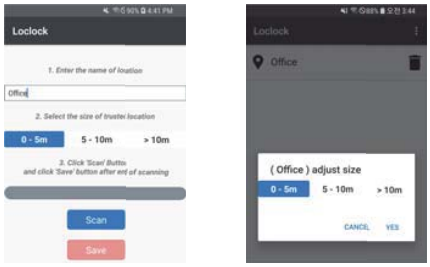
we wanted to investigate what type of trusted locations are registered in the real world, and gauge how useful the application is in reducing users' manual phone unlock burden.

Some of the public locations mentioned by the participants like cafes or gyms seem to be insecure with respect to preventing unauthorized phone access. It was our objective to analyze security implications of allowing users to freely register unlimited number of trusted locations, and select different location coverage sizes. By implementing an application based on the identified requirements (see Section 3.3), we also wanted to validate the relevance of those requirements. The ethical perspective of the field study was validated through an IRB at a university.

5.1 Methodology

We recruited 30 participants who are aged 18 years or older, and own a phone with Android 8.0 or below[1]. However, one participant dropped out on the second day of the study. Therefore, we performed our analyses on the 29 participants who completed the study. We posted advertisements for recruitment on online notice boards at a university and selectively invited people from local communities based on their age and work experiences to ensure that the overall demographic proportions are similar to those presented in [2]. To achieve strong ecological validity, we asked the participants to install our Loclock Android application (described in Section 4) on their own phone, and use it for 3 weeks. The participants were compensated for their time with a USD 200 gift card. All user interactions with Loclock (e.g., registering trusted locations, location size adjustments), WiFi data, GPS data, phone lock, and unlock events were logged. To comply with the ethical expectations of IRB, we collected all the data, removed their personally identifiable information, and stored them in an encrypted database. Moreover, only the three researchers who were approved by the IRB committee had access to the data.

Before starting the study, the participants were informed about the purposes of the study, provided with instructions, and asked to sign a consent form. They were also informed that participation is voluntary and confidential, and they can terminate the study without penalty. We asked them to submit their demographics information and install Loclock on their phones. We explained that their phones would be automatically unlocked when they move to registered trusted locations. We asked participants to turn off their current lock options for the study and switch to using the lock options provided by Loclock during the 3-week study period. We informed the participants that the only change in phone security configuration is that biometric lock mechanisms will not be available and that PIN/patterns/passwords can be used the same way. We then explained how trusted locations could be registered, removed, and modified (size changes). To ensure that the participants fully understood how Loclock works, we thoroughly explained all the features available and asked them to explain those features again. We also explained how an explicit unlock method, PIN, pattern, or password, can be registered on Loclock[2]. Loclock automatically locks a user's phone when the user carries it far away from a registered trusted location; the user should then use an explicit unlock method to unlock the phone. The setup screen of Loclock is illustrated in Figure 3. Users can freely remove or adjust the size of a registered trusted location.



(a) Location registration (b) Size adjustment

Figure 3: Loclock setup screen.

Participants were instructed to register and remove trusted locations freely, and select and adjust trusted location sizes based on their needs. However, since the field study is about analyzing the participants' behaviors with respect to using location-based authentication, we asked the participants to register at least one trusted location at the beginning of the study and use it at least until the 10th day (half of the study duration) – the intention was to collect sufficient data for meaningful analysis. We explained that they could freely remove registered locations after the 10th day if they wanted to. After the 10th day, we sent out a reminder email, informing the participants that they could freely remove any of the registered locations and discontinue using Loclock. To ensure compliance, we disabled the "remove" button until the 10th day. However, to handle cases where the participants accidentally register unwanted locations, we enabled the remove button just for an hour after initial location registration and disabled it after an hour.

Finally, a closure email was sent after 3 weeks, notifying the participants to revisit and participate in a short post-interview. We first asked the participants to explain their reasons for registering trusted locations, removing registered trusted locations, and selecting location sizes. For each of the registered trusted locations, we then asked the following scenario-based question to help categorize whether a selected location is secure from unauthorized access: *"Think about who could access your phone if it was left unattended for 10 minutes in that registered location. Is there someone who*

[1]The WiFi scanning API was depreciated from Android 9.0.

[2]Loclock does not support biometric-based unlock options like fingerprints or face detection.

should not have access?" If a participant said there are individuals who should not have access, we classified it as an "insecure" place; otherwise, we classified it as a "secure" place. We then asked the participants how they feel about the ease and time taken to register trusted locations. We also asked their feelings about the overall security and usability of using Loclock to unlock their phones. A five-level Likert scale was used to answer those questions. We helped them to uninstall Loclock. At the end of the interview, we asked "*Do you want to continue using location-based authentication after the study?*"

Before conducting the field study, we performed several rounds of pilot studies with three people to fix bugs and address unclear instructions and descriptions.

5.2 Results

5.2.1 Demographics

15 out of 29 participants were female. The participants' average age was 39.4 years (σ = 12.6). 12 participants graduated high school, 7 participants had a university degree, and 6 participants had a master (or doctoral) degree. 14 different occupations were reported with "student," "secretary," and "teacher" being the top ones. To demonstrate the representativeness of our demographics, we compared the distribution of age, gender, and education information with the US smartphone population reported in [2]. The Fisher's exact tests did not show significant differences in age ($p = 0.97$), gender ($p = 0.85$), and education ($p = 0.28$). We note that the field study participants were entirely disjunct from the requirement study participants. The details of demographics are in Appendix D.

5.2.2 Registered Trusted Locations

To satisfy the field study objectives described above, we analyzed all trusted locations registered by participants during the entire 3 weeks. Table 3 shows the trusted locations that remained at the end of the study. Participants initially registered 43 locations on the first day and additionally registered 30 locations (see Table 4). However, the total number of registered locations finally decreased from 73 to 65 because 8 trusted locations were removed after the 10th day.

As shown in Table 3, 21 participants (72%) registered two or more locations as trusted locations. Among all participants, "home" was the most frequently registered trusted location; the second most frequently registered location was "office," and the third was "my room." These results are consistent with the findings from the interview study (see Table 1). Since "my room," "living room," "bathroom," and "kitchen" are also part of "home", "home" seems to be the most representative trusted place for location-based authentication.

Interestingly, 6 participants registered "church" as a trusted location. Although the numbers were small, some participants also registered other public locations such as "sports facility," "cafe," "library," "hospital," and "subway station entrance." These observations are also consistent with the first study results (see Table 1).

Table 3: Trusted locations remaining at the end of the study, and counts for each location type.

# Locations (# Participants)	Zero (1)	One (7)	Two (13)	Three (3)	Four (3)	Five (1)	Six (1)	**Total (29)**
Home	0	0	10	2	2	1	0	**15**
Office	0	3	5	2	2	1	0	**13**
My room	0	3	4	1	1	0	0	**9**
Church	0	0	4	2	0	0	0	**6**
Sports facility	0	0	1	1	1	2	1	**6**
Living room	0	1	2	1	1	0	0	**5**
Lecture room	0	0	0	0	1	0	2	**3**
Bathroom	0	0	0	0	2	0	0	**2**
Cafe	0	0	0	0	0	1	1	**2**
Hospital	0	0	0	0	1	0	0	**1**
Kitchen	0	0	0	0	1	0	0	**1**
Library	0	0	0	0	0	0	1	**1**
Subway station entrance	0	0	0	0	0	0	1	**1**
Total	**0**	**7**	**26**	**9**	**12**	**5**	**6**	**65**

Table 4: Numbers of trusted locations registered each day of the field study.

Day	1st	2nd	3rd	4th	5th onwards	**Total since 2nd**
Home	11	4	0	1	1	**6**
Office	8	1	3	1	0	**5**
My room	6	2	0	0	1	**3**
Church	5	1	0	1	1	**3**
Sports facility	1	3	0	1	1	**5**
Living room	7	0	0	0	0	**0**
Lecture room	1	1	1	0	0	**2**
Bathroom	0	0	0	1	1	**2**
Cafe	0	0	0	1	1	**2**
Hospital	1	0	0	0	0	**0**
Kitchen	1	1	0	0	0	**1**
Library	1	0	0	0	1	**1**
Subway station entrance	1	0	0	0	0	**0**
Total	**43**	**13**	**4**	**6**	**7**	**30**

Table 5 shows the number of trusted locations that were removed after the 10th day by each location type. Eight location types, "office," "my room," "sports facility," "lecture room," "cafe," "bathroom," "subway station entrance," and "hospital" were never removed. A common characteristic between the location types that were removed – "home," "living room," "church," "library," and "kitchen" – is that they were places that could be occupied and used by other people as well. One participant (P20) initially registered two locations but removed both of them after the 10th day. When we asked why, P20 said it was simply due to curiosity. P5 initially registered four trusted locations but removed three locations. P5 said that she registered "church" because there was just one occasion where she had to stay for an entire day, and removed it after that day. P5 also explained that she removed "living room" because "my room" was registered to cover more than 10 meters, and "my room" alone was already covering the living room area as well.

Table 5: Columns "One," "Two," and "Three" refer to the number of trusted locations that were removed after the 10th day; for example, row "Two (1)" indicates there was one participant who removed two trusted locations.

# Locations (# Participants)	One (3)	Two (1)	Three (1)	**Total (5)**
Home	0	1	1	**2**
Office	0	0	0	**0**
My room	0	0	0	**0**
Church	0	1	1	**2**
Sports facility	0	0	0	**0**
Living room	1	0	1	**2**
Lecture room	0	0	0	**0**
Bathroom	0	0	0	**0**
Cafe	0	0	0	**0**
Hospital	0	0	0	**0**
Kitchen	1	0	0	**1**
Library	1	0	0	**1**
Subway station entrance	0	0	0	**0**
Total	**3**	**2**	**3**	**8**

5.2.3 Trusted Location Sizes

Next, we analyzed the sizes of trusted locations that remained at the end of the study. Table 6 shows the number of registered sizes for each location type. "5–10m" (54%) was the most frequently selected location size, followed by "> 10m" (45%). Only one instance of "my room" was registered with a size smaller than 5 meters. Even for "my room," "5–10m" (78%) was the most preferred size. These observations indicate that regardless of location types, the participants' preferred sizes are greatly divided into "5–10m" and "> 10m."

Table 6: Numbers of remaining trusted location sizes for each location type, counted at the end of the study.

Location	0–5m	5–10m	> 10m
Home	0	6	9
Office	0	6	7
My room	1	7	1
Church	0	2	4
Sports facility	0	4	2
Living room	0	5	0
Lecture room	0	0	3
Bathroom	0	1	1
Cafe	0	2	0
Hospital	0	1	0
Kitchen	0	1	0
Library	0	0	1
Subway station entrance	0	0	1
Total	**1 (2%)**	**35 (54%)**	**29 (45%)**

Worryingly, a large portion of public locations was registered with sizes larger than 10 meters (11 out of 20 public locations), which does raise security concerns about some users' size preferences. For instance, 4 out of 6 "church" locations were registered to be larger than 10 meters in size. P8 added "subway station entrance" with the largest coverage area, explaining that he always checked the subway arrival time before entering the station and wanted the phone to be unlocked automatically at that moment. About 42% of the reasons behind size selection was a general one: "to choose a location size that sufficiently covers my daily phone usage trails." No participant mentioned security as a reason for choosing a certain location size.

5.2.4 Adjusting Trusted Location Sizes

9 participants made one attempt to change the trusted location coverage meters. Interestingly, 8 of those 9 size adjustments involved increasing the coverage meters; just one adjustment led to a decrease in coverage meter from "> 10m" to "5-10m." Figure 4 visually demonstrates size adjustments. Blue arrows show adjustments leading to size increases, and red arrows show adjustments leading to size decreases.

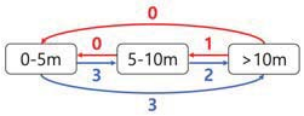
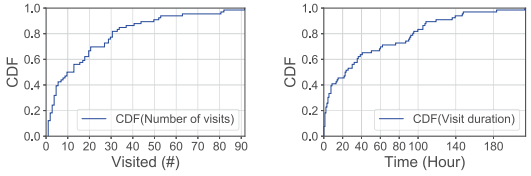


Figure 4: Size adjustments of trusted locations.

Seven participants explained that they changed location coverage size because previously chosen size was small, and did not fully cover selected locations. Two participants said that they changed location sizes just out of curiosity.

5.2.5 Visit Frequency and Duration

To examine the characteristics of trusted locations, we analyzed the number of times each trusted location was visited during the 3 weeks across all the participants, then computed cumulative distribution function (CDF) based on the number of visits for all registered trusted locations (see Figure 5(a)).

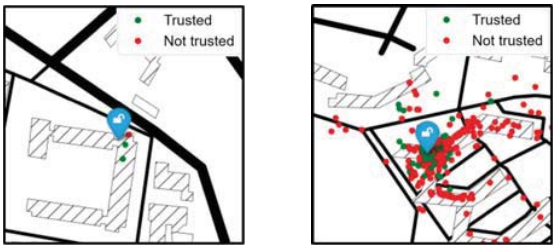


(a) Number of visits (b) Visit duration (hours)

Figure 5: CDFs computed on the total number of visits and visit duration for all registered trusted locations.

Figure 5(a) shows a significant proportion of the trusted locations were infrequently visited: over 40% of the locations were visited just 10 times or less during the 3 weeks.

We also computed CDF for the total visit duration in hours during the 3 weeks across all registered trusted locations (see Figure 5(b)). Again, it is evident that a significant proportion of the registered locations were locations where the participants did not spend much time. The participants spent 20 or fewer hours in about 45% of the registered trusted locations. These two observations indicate that some users would register places where they do not visit frequently or places where they do not necessarily spend much time.



(a) P15 (b) P2

Figure 6: Partial map view (250 by 250 meters) of where the phone was used.

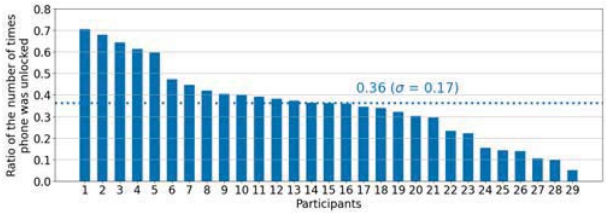


Figure 7: Distribution of the ratios in which phones were unlocked automatically through Loclock.

5.2.6 Number of Unlock Attempts

We logged the GPS and WiFi data for all locations where the participants' phone screens were turned on. We assumed that turning on the phone screen implies the intention to use it. Under this assumption, to measure the reduction of explicit unlock attempts, we counted the number of times a phone screen was turned on while the phone was unlocked by Loclock. On average, the participants tried to use their phones 44.9 times a day. This number is similar to the daily phone unlock attempts (39.9) reported in reported in [11].

Based on our assumption we counted the occurrences of `ACTION_SCREEN_ON`, which checks whether phone screens are turned on or activated. Considering that the actual number of unlocking attempts may be lower than the number of screen activation – 47.8 vs. 83.3 per day as demonstrated by [12] – the number of unlock attempts that we present may have been overestimated.

Figure 6 visualizes some locations where P2 and P15 unlocked their phones – green dots represent places where Loclock automatically unlocked phones as trusted locations, and red dots represent places where Loclock did not unlock phones. The blue unlock image represents where trusted locations were registered. Shading patterns indicate the inside of buildings. Figure 6(a) is a partial view of P15's use of the phone, showing that he or she hardly used the phone near the registered trusted location. In contrast, Figure 6(b) shows that P2 used the phone frequently near the registered trusted location. While P2 was using the phone in this area, Loclock would have automatically unlocked his or her phone many times.

Figure 7 shows the ratios of phones being unlocked automatically. The x-axis represents the participants, and the y-axis represents the ratio of the number of times a participant's phone was unlocked automatically with Loclock to the total number of unlock attempts. On average, Loclock reduced manual unlock attempts by 36% (σ=17%) – this is shown as the dashed line in Figure 7. About 25% reduction occurred from homes, and 8% occurred from offices. 20 out of 29 participants benefited from reducing more than 30% of manual unlock attempts. The largest reduction (first participant) in manual unlock attempts was 71%: from that 71%, 49.8% auto-unlock occurred from "home," and 20.5% occurred from "my room." The smallest reduction (last participant) was just 5%: this participant did not register home or office as trusted locations.

5.2.7 Security of Registered Locations

Based on the unauthorized access scenario question and responses (see Section 5.1), we labeled a given registered location as "insecure" if a participant said her phone can be accessed by unwanted individuals (who should not have access); otherwise, we labeled it as "secure." Surprisingly, based on this labeling method, 52 out of 65 registered locations were considered insecure. Table 7 shows the number of secure and insecure locations. All public places (e.g., library or sports facility) were considered "insecure" except for one instance of church registration. Interestingly, 12 "home" were considered "insecure"; four "my room," two "living room," and two "bathroom" were also considered insecure, indicating that insider threats [8, 21] may exist. Even after learning that those 52 locations are exposed to potential unauthorized access, the participants wanted to continue using 45 of them to automatically unlock phones mainly due to "phone unlock convenience" (31) or "low (perceived) likelihood of phones being attacked" (14). As for the 13 locations considered secure, 11 of them were "home" related locations.

Table 7: Counts for secure and insecure locations.

# Locations	Secure	Insecure	**Total**
Home	3	12	**15**
Office	1	12	**13**
My room	5	4	**9**
Church	1	5	**6**
Sports facility	0	6	**6**
Living room	3	2	**5**
Lecture room	0	3	**3**
Bathroom	0	2	**2**
Cafe	0	2	**2**
Hospital	0	1	**1**
Kitchen	0	1	**1**
Library	0	1	**1**
Subway station entrance	0	1	**1**
Total	**13**	**52**	**65**

As for the location coverage sizes, regardless of whether locations are considered secure or insecure, the participants

preferred selecting sizes larger than 5 meters in radius ($p = 1.00$, Fisher's exact test). These results are summarized in Table 8.

Table 8: Secure and insecure locations and coverage sizes.

Location	0–5m	5–10m	> 10m	**Total**
Secure	2	6	5	**13**
Insecure	8	23	21	**52**

5.2.8 Post Study Survey Results

Location registration difficulty. As part of the post study survey, we asked the participants about their feelings toward the easiness of registering a trusted location. The participants' responses are summarized in Figure 8. About 86% felt that it was easy to register trusted locations, and there was no participant who felt it was difficult.

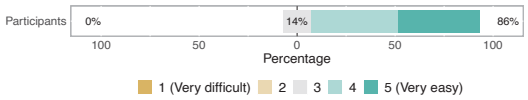


Figure 8: Easiness of registering trusted locations.

Time taken to register trusted locations. We also asked how the participants felt about the time it took for them to register trusted locations. Note, the time taken to collect and store WiFi RSSI values is one minute. Their responses are summarized in Figure 9. About 48% of the participants felt that the time taken to register trusted locations was fast. 21% felt that it was slow.

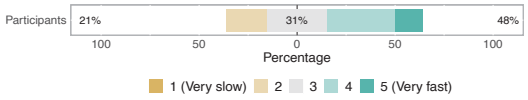


Figure 9: Fastness of registering trusted locations.

Security of Loclock. We asked how the participants felt about the security offered by location-based authentication; their responses are summarized in Figure 10. About 62% of the participants felt that using Loclock was secure; only 7% felt that it was insecure. The low reported FARs (1% on average) are one explanation as to why the participants may have felt that Loclock was secure to use.

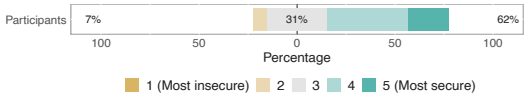


Figure 10: Security of using Loclock.

Convenience of Loclock. We also asked how the participants feel about the convenience associated using Loclock to automatically unlock their phones. Their responses are summarized in Figure 11. About 59% of the participants felt that Loclock was convenient to use. The common reason was because of its automatic unlock capabilities. P15 mentioned that he wants to continue using Loclock even after the study. 10 participants felt that it was inconvenient. 7 of those 10 had to deal with unintended termination of Loclock due to insufficient memory or communication errors at some point during the study, and mentioned this as the main reason. Two participants mentioned "no support for fingerprint scanner." Only one participant mentioned battery drain as the reason.

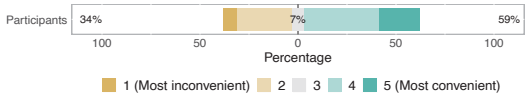


Figure 11: Convenience of using Loclock.

Intentions for future use. Finally, the participants were asked whether they would continue to use Loclock. 22 out of 28 participants said "yes," indicating "phone unlock convenience" as the main reason. As for the 6 participants who said "no," the main reason for not willing to use it in the future was "concerns about information leakage."

5.3 Limitations

We made it mandatory to register at least one trusted location and use it for the first 10 days. Also, Loclock does not support biometric-based authentication options. These constraints may have affected the ecological validity of the field study. To study the effects of having previous biometric authentication experience, we divided the field study participants into two groups: 16 participants who were using at least one biometric scheme prior to the study and 12 participants who were not using any. We analyzed the statistical differences between the two groups. As for the number of registered locations, we did not find any significant difference between them ($p = 0.82$, Mann-Whitney U test). However, as for the location coverage sizes, we did find a significant difference between the two ($p < 0.001$, Fisher's exact test). One possible explanation is that those who were previously using biometric schemes tried to minimize the burden of using passwords (for the study) by selecting large location coverage areas.

Loclock was not optimized for location detection accuracy and battery use. Also, its GUI was not optimized for usability. All of these limitations may have affected the way the participants felt about the overall security and usability of Loclock.

As explained above, while measuring the benefits of reducing the number of explicit unlock attempts we counted the number of times phone screen was activated – we could have overestimated the manual unlock benefits due to this limitation.

6 Discussions

6.1 Security Concerns

The results from the interviews and field studies raise two important security concerns: (1) people tend to add a variety of insecure locations (52 out of 65 registered locations were considered insecure by our definition) and are willing to continue to use them even after becoming aware of potential unauthorized phone access threats, and (2) a significant proportion of such insecure locations are added with the largest coverage areas (larger than 10 meters). Moreover, some participants added locations where they spend a small amount of time as trusted locations – many of them being public places exposed to phone theft. All of those observations indicate that location-based authentication schemes could expose new security threats that adversaries may exploit. For instance, if an adversary has some information about a victim's location history, the adversary could try to steal the victim's phone, go near a pre-registered trusted location, and access the phone contents without having to guess PIN or pattern. An insider could try to access phone contents when the victim leaves the phone unattended inside a registered trusted location. Such threats could compromise the entire phone security and need to be mitigated carefully.

To mitigate them, location-based authentication systems need to be designed to help users adequately understand the security risks associated with adding certain locations or choosing large sizes. For instance, we could ask a similar phone access scenario question (see Section 5.1) while adding a new location, and help users become aware of any unwanted access that might occur. Current Smart Lock implementation provides a simple guide for users to add their homes as trusted locations ("Keep device unlocked at Home") without informing users about the possibility of insider threats. Again, security risks related to insider threats need to be conveyed before offering recommendations to add homes.

However, such mitigation strategies might not be sufficient (as observed from Section 5.2.7) if users still select and use insecure locations, thinking that threat likelihood is low or focusing merely on the usability benefits. Therefore, we believe more protective measures need to be deployed with a location-based authentication scheme: for instance, one could design it so that phones must first be unlocked with an explicit unlock scheme – it would then stay unlocked within a detected trusted location. Since most usability benefits came from homes and offices, another security measure could disallow the registration of any other location. Infrequently visited locations could be deleted automatically after notifying users.

6.2 Usability

Our field study results show that the participants were willing to continue to use Loclock. As described in Section 5.1, after the 10th day, we informed the participants that they could freely remove all registered locations and use manual unlock instead. Just one participant (out of 29) stopped using Loclock after the 10th day. Table 4 and 5 show that the number of registered locations increased from 43 to 65 during the 3-week period. Through the use of Loclock, the participants managed to reduce about 36% of manual unlock attempts (mostly used at homes or offices) – demonstrating clear usability benefits (and usefulness) of location-based authentication schemes.

As shown in Section 5.2.8, 21% of the participants felt that the trusted location registration process was slow. The current Loclock implementation required the participants to wait for a minute to collect WiFi RSSI values but the entire one minute data might not be necessary to maintain the reported accuracy. Future design should consider shortening this setup time (e.g., to 30 seconds) while trying to maintain similar level of detection accuracy.

One participant mentioned the battery drain issue and said Loclock was inconvenient to use because of its heavy battery usage. Although the background logging services contributed to more battery being used, overall, its battery use was far greater than the tolerable levels mentioned in the requirements. Since continuous WiFi sensing is a battery-intensive operation, future work should look at other possible indicators that would help identify a physical location and use less battery; e.g., detecting the presence of known (previously paired) Bluetooth devices.

Even though the reported FARs and FRRs were small, we imagine that real-world error rates may be higher. A recent study [23] demonstrates that it is important to provide a well-designed user-in-the-loop user experience so that users can manually deal with inaccuracies. Following their design guidelines, we may give users the ability to adjust the threshold based on their preferences to reduce error rates.

7 Conclusion

Through interviews and a field study, we identified essential requirements for building usable and secure location-based authentication services: users prefer to register fine-grained indoor locations and adjust location coverage sizes. Using a location-based authentication application, the participants, on average, were able to reduce 36% of explicit authentication attempts, demonstrating clear usability benefits. Most of the participants continued using the automatic unlock feature despite being informed that they could stop using it and return to manual unlocks. However, the field study findings also revealed that people tend to register insecure locations due to convenience or perceived low likelihood of phones being attacked in those locations. Even after being informed about potential phone access threats, most of the participants said they would continue using insecure locations. Such risks would probably exist in commercialized services like Smart Lock, and need to be mitigated.

Acknowledgments

This work was supported by Samsung Research, NRFK (2019R1C1C1007118) and IITP (2019-0-01343). The authors would like to thank all the anonymous reviewers. Note that Hyoungshick Kim is the corresponding author.

References

[1] Choose when your Android device can stay unlocked. https://support.google.com/android/answer/9075927?hl=en.

[2] Mobile Technology and Home Broadband 2019. https://www.pewresearch.org/internet/2019/06/13/mobile-technology-and-home-broadband-2019/.

[3] M. Abbas, M. Elhamshary, H. Rizk, M. Torki, and M. Youssef. WiDeep: WiFi-based Accurate and Robust Indoor Localization System using Deep Learning. In *Proceedings of the 17th International Conference on Pervasive Computing and Communications*, 2019.

[4] Ioannis Agadakos, Per Hallgren, Dimitrios Damopoulos, Andrei Sabelfeld, and Georgios Portokalidis. Location-Enhanced Authentication Using the IoT: Because You Cannot Be in Two Places at Once. In *Proceedings of the 32nd Annual Conference on Computer Security Applications*, pages 251–264, 2016.

[5] Jiayi Chen, Urs Hengartner, Hassan Khan, and Mohammad Mannan. Chaperone: Real-time Locking and Loss Prevention for Smartphones. In *Proceedings of the 29th USENIX Security Symposium*, pages 325–342, 2020.

[6] P. Davidson and R. Piché. A Survey of Selected Indoor Positioning Methods for Smartphones. *IEEE Communications Surveys Tutorials*, 19(2):1347–1370, 2017.

[7] Dorothy E Denning and Peter F MacDoran. Location-based authentication: Grounding cyberspace for better security. *Computer Fraud & Security*, 1996(2):12–16, 1996.

[8] Serge Egelman, Sakshi Jain, Rebecca S. Portnoff, Kerwell Liao, Sunny Consolvo, and David Wagner. Are You Ready to Lock? In *Proceedings of the 21st ACM SIGSAC Conference on Computer and Communications Security*, page 750–761, 2014.

[9] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.

[10] L. Fridman, S. Weber, R. Greenstadt, and M. Kam. Active Authentication on Mobile Devices via Stylometry, Application Usage, Web Browsing, and GPS Location. *IEEE Systems Journal*, 11(2):513–521, 2017.

[11] Marian Harbach, Alexander De Luca, and Serge Egelman. The Anatomy of Smartphone Unlocking: A Field Study of Android Lock Screens. In *Proceedings of the 34th Conference on Human Factors in Computing Systems*, pages 4806–4817, 2016.

[12] Marian Harbach, Emanuel von Zezschwitz, Andreas Fichtner, Alexander De Luca, and Matthew Smith. It's a Hard Lock Life: A Field Study of Smartphone (Un) Locking Behavior and Risk Perception. In *Proceedings of the 10th Symposium On Usable Privacy and Security*, pages 213–230, 2014.

[13] Sebastian Hilsenbeck, Dmytro Bobkov, Georg Schroth, Robert Huitl, and Eckehard Steinbach. Graph-based Data Fusion of Pedometer and WiFi Measurements for Mobile Indoor Positioning. In *Proceedings of the 14th ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 147–158, 2014.

[14] Daniel Hintze, Eckhard Koch, Sebastian Scholz, and Rene Mayrhofer. Location-Based Risk Assessment for Mobile Authentication. In *Proceedings of the 7th International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, page 85–88, 2016.

[15] Fudong Li, Nathan Clarke, Maria Papadaki, and Paul Dowland. Active authentication for mobile devices utilising behaviour profiling. *International Journal of Information Security*, 13(3):229–244, 2014.

[16] Tao Li, Yimin Chen, Jingchao Sun, Xiaocong Jin, and Yanchao Zhang. ILock: Immediate and Automatic Locking of Mobile Devices against Data Theft. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, page 933–944, 2016.

[17] Kathleen Macqueen, Eleanor McLellan-Lemal, K. Bartholow, and B. Milstein. Team-based codebook development: Structure, process, and agreement. *Handbook for team-based qualitative research*, pages 119–135, 2008.

[18] Upal Mahbub and Rama Chellappa. PATH: Person authentication using trace histories. In *Proceedings of the 7th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference*, pages 1–8, 2016.

[19] Claudio Marforio, Nikolaos Karapanos, Claudio Soriente, Kari Kostiainen, and Srdjan Capkun. Smartphones as Practical and Secure Location Verification Tokens for Payments. In *Proceedings of the Network and Distributed System Security Symposium*, pages 23–26, 2014.

[20] Masoud Mehrabi Koushki, Borke Obada-Obieh, Jun Ho Huh, and Konstantin Beznosov. Is Implicit Authentication on Smartphones Really Popular? On Android Users' Perception of "Smart Lock for Android". In *Proceedings of the 22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 1–17, 2020.

[21] Ildar Muslukhov, Yazan Boshmaf, Cynthia Kuo, Jonathan Lester, and Konstantin Beznosov. Know Your Enemy: The Risk of Unauthorized Access in Smartphones by Insiders. In *Proceedings of the 15th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 271–280, 2013.

[22] F. S. Park, C. Gangakhedkar, and P. Traynor. Leveraging Cellular Infrastructure to Improve Fraud Prevention. In *Proceedings of the 25th Annual Computer Security Applications Conference*, pages 350–359, 2009.

[23] Quentin Roy, Futian Zhang, and Daniel Vogel. Automation Accuracy Is Good, but High Controllability May Be Better. In *Proceedings of the 37th ACM Conference on Human Factors in Computing Systems*, 2019.

[24] Johnny Saldaña. *The coding manual for qualitative researchers*. Sage, 2015.

[25] Y. Shu, C. Bo, G. Shen, C. Zhao, L. Li, and F. Zhao. Magicol: Indoor Localization Using Pervasive Magnetic Field and Op-

portunistic WiFi Sensing. *IEEE Journal on Selected Areas in Communications*, 33(7):1443–1457, 2015.

[26] Feng Zhang, Aron Kondoro, and Sead Muftic. Location-Based Authentication and Authorization Using Smart Phones. In *Proceedings of the 11th International Conference on Trust, Security and Privacy in Computing and Communications*, pages 1285–1292, 2012.

A Demographics in the Requirement Study

Table 9 presents the demographics of the participants in the requirement study.

Table 9: The demographics of the requirement study.

Gender		
Female	10	(55.6%)
Male	8	(44.4%)
Age		
19–24	2	(11.1%)
25–34	5	(27.8%)
35–44	6	(33.3%)
45–54	2	(11.1%)
55–64	3	(16.7%)
Education		
Less than high school	0	(0.0%)
High school	3	(16.7%)
Professional School	0	(0.0%)
University (Bachelor's)	9	(50.0%)
Master of PhD	6	(33.3%)
Other	0	(0.0%)
Occupation		
Managers	3	(16.7%)
Professionals	2	(11.1%)
Clerical Support Workers	4	(22.2%)
Service and Sales Workers	3	(16.7%)
Craft and Trades Workers	1	(5.6%)
Machine Operators	1	(5.6%)
Elementary Occupations	0	(0.0%)
Students	3	(16.7%)
Self-employed	0	(0.0%)
Unemployed/Retired/Disabled	1	(5.6%)

B Tolerable battery consumption

Table 10 shows the distribution of participants' responses in the requirement study. We can see that tolerable battery usage percentage mainly ranged from 5 to 15%.

Table 10: Tolerable daily battery usage levels.

Battery usage	5–10%	10–15%	15–20%	20–25%	**Total**
Frequency	9	6	1	2	**18**

C Lock/Unlock Failure Rate Evaluation

C.1 Methodology

Using the Loclock application installed on a Samsung Galaxy S8 phone, we collected WiFi RSSI values from 3 locations. For each location, we created a grid layout with one meter spacing between two grid points, covering the entire floor space. At every grid point, we collected RSSI values for one minute. The first data collection took place at a single floor in a small office building (L1) – its size is 46 by 10 meters; the number of collected BSSIDs ranged from 100 to 120. Similarly, the second location was a single floor in another office building (L2) – its size is 55 by 20 meters; the number of collected BSSIDs ranged from 15 to 20. The last location was a university laboratory (L3) that consists of 14 computer desks – its size is 11 by 7 meters; the number of collected BSSIDs ranged from 60 to 80.

After creating meter-by-meter RSSI maps for the three locations, respectively, we physically moved to a *central* position in the grid for each location, and registered that central spot as a trusted location starting point using Loclock. WiFi RSSI values, collected for a minute, were then used to compute the pre-stored trusted location RSSI vector. Using the meter-by-meter RSSI maps and pre-stored trusted location RSSI vectors, we measured unlock failure and lock failure rates for different trusted location coverage areas.

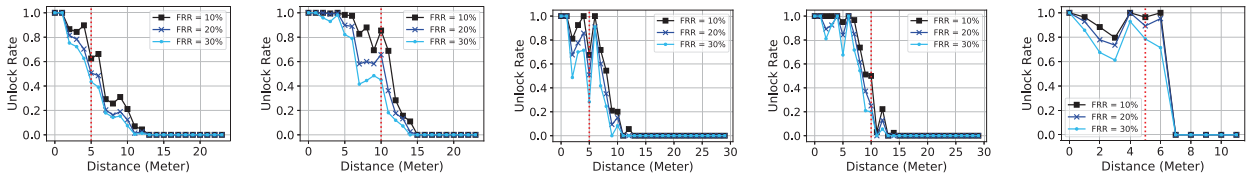
C.2 Evaluation Results

We measured lock and unlock failure rates of Loclock. Lock failure rates represent "false acceptance rates" (FAR) that measure the error rates reflecting the number of times a phone accidentally unlocks itself when a user is not inside a trusted location coverage area. This error rate is associated with the security of Loclock since the user's phone would be unlocked automatically in unknown (potentially untrusted) environments. Unlock failure rates represent "false rejection rates" (FRR), measuring the error rates for when a phone does not unlock automatically when a user has physically moved to a trusted location coverage area. This error rate would affect the usability of Loclock since users would have to unlock their phones manually.

Table 11: Lock and unlock failure rates of Loclock.

Coverage		5m			10m		
FRR		10%	20%	30%	10%	20%	30%
FAR	L1	20.0%	13.8%	11.3%	23.0%	14.9%	9.1%
	L2	13.2%	9.8%	6.4%	3.8%	1.8%	1.2%
	L3	20.9%	19.6%	16.1%	-	-	-
HTER	L1	15.0%	16.9%	20.7%	16.5%	17.5%	19.6%
	L2	11.6%	14.9%	18.2%	6.9%	10.9%	15.6%
	L3	15.5%	19.8%	23.1%	-	-	-

For the two locations (L1) and (L2), we measured FRRs and FARs for two trusted location coverage sizes: one with a circular coverage radius of 5 meters and another with a coverage radius of 10 meters. As for the third location (L3), the university laboratory, we only evaluated error rates for 5 meter radius coverage because its size is 11 by 7 meters. For each coverage area, we measured three sets for FRR and FAR, fixing FRRs to 10, 20, and 30% – this would give us three specific RSSI threshold values that guarantee those three FRR



(a) L1 (radius of 5 meters) (b) L1 (radius of 10 meters) (c) L2 (radius of 5 meters) (d) L2 (radius of 10 meters) (e) L3 (radius of 5 meters)

Figure 12: Measuring phone unlock rates with varying trusted location coverage areas (5 and 10 meters) in small office building (L1), large office building (L2) and small university laboratory (L3).

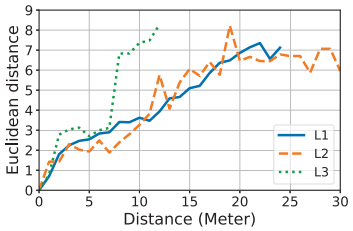


Figure 13: Changes in Euclidean distance while moving away from the originally registered spots in each of the three locations.

rates – and measuring resulting three FARs based on the three threshold values. These FRR and FAR results are summarized in Table 11. As the results show, at both FRR 10 and 20% threshold values, the FARs were contained around 20% (except for L2 that went as high as 23%). The half total error rates (HTER), computed by averaging FARs and FRRs, are all below 20% when FRRs are fixed at 10 and 20%. Referring back to the "unlock/lock failures" requirement (willing to tolerate one or two out of 10 failures), these FRR/FAR results indicate the next field study participants would likely experience reasonable and tolerable error rates. Further, Figure 12 shows the phone unlock rates in L1, L2, and L3, measuring the number of times the phone would be unlocked within the radius meters shown in the x-axis. The dotted vertical red lines show the coverage radius, 5 and 10 meters, respectively. We note that the change in WiFi RSSI values is not only determined by physical distances between access points and a user's phone; there are other factors such as physical barriers between phone and access points – the unlock rate results do not always decrease linearly based on varying distances (moving away from registered spots), and guaranteeing meter-level accuracy with just RSSI values would be infeasible. Figure 13 shows how the Euclidean distance (ED) values change with varying distances for each of the three locations. Each of the three lines in the graph represent the three different locations, and how ED changes differently based on their physical characteristics. As for L3, the sudden jump in ED is caused by walking out the laboratory door.

D Demographics in the Field Study

Table 12 presents the demographics of the participants in the field study.

Table 12: The demographics of the field study.

Gender		
Female	15	(51.7%)
Male	14	(48.3%)
Age		
19–24	4	(13.8%)
25–34	7	(24.1%)
35–44	6	(20.7%)
45–54	8	(27.6%)
55–64	4	(13.8%)
Education		
Less than high school	0	(0.0%)
High school	13	(44.8%)
Professional School	3	(10.3%)
University (Bachelor's)	7	(24.2%)
Master of PhD	6	(20.7%)
Other	0	(0.0%)
Occupation		
Managers	0	(0.0%)
Professionals	3	(10.3%)
Clerical Support Workers	8	(27.6%)
Service and Sales Workers	2	(6.9%)
Craft and Trades Workers	0	(0.0%)
Machine Operators	0	(0.0%)
Elementary Occupations	0	(0.0%)
Students	10	(34.5%)
Self-employed	2	(6.9%)
Unemployed/Retired/Disabled	4	(13.8%)
Current unlock methods		
Password	4	(8.9%)
Pattern	22	(48.9%)
PIN	2	(4.4%)
Finger	15	(33.4%)
Face	1	(2.2%)
Knock Code	1	(2.2%)

Please do not use !?_ or your License Plate Number: Analyzing Password Policies in German Companies

Eva Gerlitz*
Fraunhofer FKIE

Maximilian Häring*
University of Bonn

Matthew Smith
University of Bonn, Fraunhofer FKIE

Abstract

Password composition policies (PCPs) set rules that are intended to increase the security of user-chosen passwords. We conducted an online survey and investigated the employee-facing authentication methods of 83 German companies and the extracted 64 PCPs. We compared the password policies to recommendations proposed by institutions and related work. We found that many companies still require several character classes to be used as well as mandating regular password changes. Short and complex passwords are more often enforced than alternative mechanisms, such as minimum-strength requirements, that related work found more usable. Many of the policies were in line with recommendations given through the German Federal Office for Information Security (BSI). At the same time, there is high heterogeneity in the reported elements. Based on a selection of the main elements (password age, complexity, minimal length), at most seven out of the 64 PCPs are identical. The company size does not seem to play a significant role in the configuration of the PCPs.

1 Introduction

Passwords as a security measure are the daily reality of users working with computers, and even with technologies like FIDO2, they will likely stay for a while. It is well known that users sometimes choose weak passwords regarding their security effect. Websites and companies thus try to prevent this by using password composition policies (PCPs). These policies constrain the passwords users can choose, e.g., by preventing commonly chosen passwords. However, poorly chosen PCPs can be detrimental to usability and security [34]. A large body of work looks at PCPs in end user-facing websites, e.g., [15, 30, 31, 38], and how users cope with PCPs, e.g., [23, 26, 29, 33]. In this paper, we look at this topic from the view of those who manage PCPs. We conducted a survey with IT staff from 83 German companies. We focused on employee-facing PCPs since their passwords often protect accounts of great value for hackers (e.g., espionage or access to large amounts of user data).

To help companies with the creation of PCPs, organizations like the American NIST (National Institute of Standards and Technology) [22], OWASP (Open Web Application Security Project) [2] or the German BSI (Bundesamt für Sicherheit in der Informationstechnik, the German Federal Office for Information Security) [6] provide guidelines. To analyze if and how these guidelines affect the creation of PCPs, we surveyed what PCPs our participants used for company-wide user accounts or company email accounts and what information sources they used during the creation of the PCPs. We also surveyed what their experiences and perceptions of the PCPs is. We found a very heterogeneous set of PCPs with a surprising number of creative and unique PCP elements.

In this paper, we

- give an overview of the PCP landscape of 83 German companies.
- look at possible influences on and of PCPs.
- compare the identified PCP elements with recommendations.

The rest of the paper is structured as follows: First, we give an overview of relevant related work regarding PCPs and their effects on the resulting passwords, then we describe the methodology of the study, followed by the results, discussion, and directions for future work.

* These authors contributed equally to this work.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

2 Related Work

There is a large body of literature on various aspects of password authentication. In the following section, we discuss previous work that is most relevant to our study. First, we give a short overview of the analysis of existing PCPs on websites, followed by the effects of different PCPs on end user behavior. Finally, we give a summary of guidelines given by organizations like NIST or the BSI.

2.1 Analysis of Existing Password Composition Policies

In 2010, Florêncio et al. [21] examined the password policies of 75 websites, including top, high and medium traffic sites as well as banks, universities and government sites. They calculated the minimum strength of each password policy using the cardinality of the minimum character set required and the minimum length given in the policy. Afterward, they analyzed if different characteristics correlate with stronger password policies but found no correlation between the website's size, the number of users, or the frequency of attacks. Instead, they noted a strong inverse correlation between password policy strength and sites that accept advertising and sponsored links. The authors hypothesized the necessity of those websites to have high usability to keep users on their site.

Mayer et al. [28] replicated and extended this study in 2016 by analyzing the password policies of the same websites as visited by Florêncio et al. and additionally investigating a corresponding sample of German websites. They noted that the average strength of the password policies had grown significantly in the US, except for websites that display third-party advertisements. In all samples, inverse correlation was found for users visiting a website with a clear competitor regarding their service. While comparing the password policies of German websites and those from the US, the authors noted a much smaller median of policy strength on German websites, with especially weak policies on banking websites.[1]

2.2 Effect of Password Policies on Usability and Password Strength

Komanduri et al. [27] studied password strength and user sentiment across four password composition policies in 2011. For this, they invited 5000 people to participate in an online study where participants had to create a password that they had to recall two days later. The policies requested a certain length (8 vs. 16 characters) either alone, with an additional complexity requirement, or the non-existence of dictionary words in the chosen password. They found that participants across all conditions used at least 2.2 digits, while symbols mainly were used if a policy requested to do so. Also, requiring a high complexity led to passwords with a higher entropy than other policies. At the same time, high complexity and the ban of dictionary words made password creation more complicated, with only 17.7% of participants being able to create a password in one try compared to the 52 to 84 % with other policies. The authors noted a correlation between storing passwords and the use of higher-entropy passwords. Of the four tested password composition policies, the one asking for at least 16 characters but not requiring anything else seemed to be the best trade-off between usability and security of the resulting passwords.

The same approach was followed by Kelley et al. [26] in 2012, Shay et al. in 2014 [33] and 2016 [34] and Tan et al. [35] in 2019. Kelley et al. [26] tested the effect of policies of different lengths and complexities as well as the presence of password blocklists, which varied in their size and complexity. They found that larger and more complex blocklists lead to stronger passwords. Shay et al. [34] examined 15 password policies by inviting 20,000 participants. Their password composition policies included policies that required only a minimal length or a length in combination with complexity or a certain number of words. The authors found that requiring a longer password with less complexity made it easier for participants to create and recall them while being less likely to be guessed. While experimenting with password blocklists, they noted that substring blocklists made passwords more secure without making recalling them more difficult. Policies that only requested minimal lengths were found to be usable; however, many of the resulting passwords were very weak. Additionally, the authors found the frequently used PCP consisting of a minimum of 8 characters and one character of each character class to be less usable and secure than some other tested policies. Based on their findings, the authors also gave recommendations for service providers regarding password composition policies.

Tan et al. [35] tested 21 policies, including composition requirements, blocklist requirements (using different lists and four different matching algorithms) and minimum-strength requirements (i.e., the number of guesses needed). During creation, a password meter showed compliance with the requirements and gave additional hints on how to increase the security once compliance was met. The study was completed by 6477 participants. The authors found that character-class requirements are annoying while simultaneously resulting in passwords that can be easily cracked using state-of-the-art password-cracking tools. They additionally saw usability differences when comparing different blocklists and found no benefit of requiring four character classes in addition to a large blocklist. They recommend using a minimum-strength requirement in combination with a length requirement and rate the benefit of minimum-strength higher than blocklists in protecting against offline attacks.

Ur et al. [37] asked 49 participants to create an account for

[1] It should be noted, though, that the login for costumers is protected by rate-limiting. Further actions need to be approved by a second factor [8].

fictitious banking, email, and news websites while thinking aloud to understand common password patterns and users' misconceptions about password strength. The authors found that while some weak passwords were created consciously, most were a result of misconceptions, e.g., that a "!" at the end makes a password more secure or that hard to spell words are securer than easy ones. Additionally, many participants demonstrated misconceptions regarding possible attacks, believing that personal data as passwords is secure as long as it is not known publicly. When confronted with policies that required the participants to add numbers or symbols to their password, which they had not included before, many simply appended one to their password.

Inglesant et al. [25] let 32 staff members of two different companies keep a password diary for one week and interviewed them regarding the details of each password. They found that the policies existing in 2010 were too complex, which, in the worst case, harmed the (organizational) productivity.

In a study of passwords collected from over 25.000 members of their university, Mazurek et al. [29] found the passwords of people who were annoyed by the complex password composition policy to be weaker.

In 2010, Zhang et al. [39] examined whether password expiration meets its intended purpose. For this, they analyzed a data set consisting of 7700 accounts. They found that 41% of the new passwords can be broken with knowledge about previous passwords for the same account within seconds, and 17% of the accounts can be broken into with five online password guesses.

Similarly, Habib et al. [24] found that 67% of the participants from an online survey self-reported creating their new password by modifying their previous one; most prominent was capitalizing a letter, which was done by 30%. Still, according to self-reports, regular password changes do not seem to lead to weaker passwords. 82% of their participants agreed that frequent password expiration secures accounts against unauthorized persons.

Using a more theoretical methodology, Shay and Bertino [32] presented an algorithm to simulate the effect of policies on security. As input, it takes details of the policy (e.g., length, per-character entropy, expiration), details of users (e.g., probability that a user remembers a seven-digit password after seeing it for the first time) and details of the service. With this, they offer administrators the possibility of testing a PCP concerning various properties of their organization.

Blocki et al. [14] presented an algorithm that takes a sample of users' preferred passwords as input. Based on this, it creates an ideal policy that maximizes the minimum entropy of the resulting distribution of passwords.

2.3 Official Recommendations

Table 2 (Appendix) shows the recommendations for password policies given by the American NIST [22] and the German BSI [5]. A few months after we conducted this study, the BSI changed their recommendations substantially in the area of PCPs. We will thus refer to their recommendations that were present during our study as "old", and the revised BSI recommendations as "new".

NIST published very specific recommendations, for example, regarding the minimum number of characters a password should have (minimal length), the minimum number of characters a password should be allowed to have (minimal maximal length), the number of character classes covered in a password (complexity), the time after which a password needs to be reset (maximal age), which characters should be allowed as part of the password (allowed characters) and which elements should be prohibited from usage (blocklist). On the other hand, both BSI recommendations consciously keep the recommendations vague to leave room for interpretation, for example, by stating that passwords of suitable quality should be chosen, without defining "quality". The BSI guidelines use as a basis the ISO 270012 and can be used as a help to implement it [12]. Additionally to the guideline, the BSI published implementation notes [11] with examples on how the policies could be built. The examples are based on a combination of minimal length and character classes, e.g., length of 20 to 25 and two character classes, or length of 8 to 15 and four character classes.

3 Methodology

To investigate the current state of password composition policies in German companies, we conducted a survey in late 2019 using Qualtrics [10].

The survey aimed at people who are responsible for PCPs in German companies. The questionnaire was offered in German and English since employees at this level can be from an international context.

3.1 Survey Design

Since our target audience is usually very busy and extremely hard to recruit for research purposes, we paid special attention to keeping the survey as short as possible. Thus, our survey was designed to take around ten minutes. Since the PCP is a sensitive piece of information and we were concerned that companies would not share them with us, we did not collect any information that could identify the company and any personal data from the person taking part in the study. While this would have been interesting data, we did not want to jeopardize either the company or our participants if our systems were breached.

Our survey, which can be found in Appendix A, was structured into three sections.

The first part asked whether the company uses a company-wide account per user that is centrally managed, e.g., for logging onto workstations, email, communications platforms, and the like. If this was the case, we asked for the authentication method(s) with which this account is secured, e.g., passwords, biometrics, or tokens. For those companies that did not have such central accounts, we opted to study the authentication methods for the company email accounts as we were fairly certain that most companies would have such a service. This way, we were able to include these companies and observe possible differences in these application areas.

The second section included questions regarding the authentication method(s) details, e.g., if a password composition policy is used, who created the PCP, and asked for the PCP itself. We encouraged participants to copy and paste their policy if they were allowed. Further, we asked for our participants' opinions on the authentication methods' security and usability.

The last section contained general information and demographic questions. As noted above, we collected only very minimal demographic information.

3.2 Survey Testing

Since we set ourselves a strict time limit for the survey, it proved challenging to formulate short enough questions to not slow down the survey but also unambiguous enough to gather useful data. The survey underwent five internal iterations, and we conducted a pilot study with the VP of Security of a large multi-national company and an administrator responsible for a small organization. We integrated the feedback from the pilot study into the final version of the survey.

3.3 Recruitment

We recruited our participants through several channels. The most effective channel by far was a newsletter sent by the BSI ($n = 69$ valid data sets of 83 valid data sets in total). We also recruited via contacts of two german digital associations (Bitkom [1] and Cyber Security Cluster Bonn [4]) and personal contacts.

Our survey was targeted at the person within the company responsible for the authentication system and the PCP. Since we had no way of contacting them directly, we clearly stated that only these people could fill out the survey and requested that the survey link be passed on to this person within the company. As we offered opt-out options, we believe that an accidental non-decision maker would have to have had malicious intent to affect the results negatively. In total, 110 participants took part. The participants were not compensated for their time.

3.4 Data Quality

Since we expected a heterogeneous set of PCPs and asked questions that are either sensitive or broad and thus may not apply to every participant, we included "Other", "I do not wish to make a statement", and "I do not know" options to questions (see also Section 3.6 and Section 6). This way, we wanted to prevent that the participants leave after facing a question that they could or did not want to answer.

Before analyzing the data, we checked for duplicate companies by using the company demographics and policies. We saw nothing to suggest that one company participated more than once. We also manually went through all the complete answers and excluded one participant who gave answers in the open texts, which led to the conclusion that they had not understood the previous questions. We also excluded 25 participants who had a completion rate lower than 50%. Most (21) of them closed the survey after answering whether there is a centrally managed account and, if so, what authentication methods can be used to log in.

In the end, we were left with 83 complete, valid data sets, and 77 policies.

3.5 Data Analysis

In our analysis, we separated the 77 password policies by their usage for a centrally managed account ($n = 64$, in the following called PWA for "Password Account") and those applying for email accounts ($n = 13$).

To analyze the PCPs, we used open coding as described by Corbin et al. [19]. Even though the policies mainly were enumerations of several elements and did not allow much room for interpretation (with few exceptions like "no *easy* passwords"), two researchers independently coded the policies to reduce errors. As suggested by Campbell et al. [16], we developed a code book by separately coding a small set of policies (n = 10) and comparing the codes. This then served as a base for future codes. After coding all the remaining policies independently, the codes were compared, and the inter-coder agreement was calculated using Cohen's kappa coefficient (κ) [18]. Our agreement was 81.48. A value above 0.75 is considered a good level of coding agreement [20]. We were able to resolve all conflicts.

We found a large set of possible properties concerning a PCP which we used to categorize each PCP based on its attributes (e.g., minimal length = 8 characters, minimal age = 1 day, maximal age = 90 days). We opted against calculating the strength of the PCPs as done by Florêncio et al. [21], and Mayer et al. [28], which only describes the theoretical size of the possibilities. It is also acknowledged by Florêncio et al. and Mayer et al. themselves that this is not a good metric to calculate the resulting password strength.

However, we discuss compliance with recommendations regarding PCPs from related work.

This study contributes data with an exploratory approach guiding to further research themes. Trends and interesting data were only very rarely tested on statistical significance to reduce the problems of multiple comparisons analysis.

When looking for statistical significance, we corrected the results with the Bonferroni–Holm method, also taking tests into account that we did but do not report. In the following sections, the stated p is that after correction. Percentages are reported rounded.

3.6 Ethics

The companies were asked details of their authentication methods and policies, which could give indications of vulnerabilities. To keep risks of exposure low, we did not collect any information that could identify a company or individual.

If participants included their company's name in one of the free text fields, we anonymized the answer before analyzing it. Additionally, the respondents were given an explicit option to answer questions with "I do not want to answer". Before the survey began, there was an introduction to the study, and participants had to consent.

The Research Ethics Board of our university reviewed and approved our study.

4 Results

In the following section, we will present the results of the survey. First, we present the demographics and the authentication methods used by the participating companies. This is followed by an analysis of the present password composition policies and their different components that respect to Table 2. We conclude this section with an overview of the potential impacts different authentication methods have.

4.1 Demographics

Table 3 (Appendix) shows the size (number of employees) of the participating companies ($n = 83$) and the number of desktop clients the participants had to handle.

We asked the participants what situations regarding their emails apply to their companies. 60 (72%) stated that employees can access their emails outside the company network. In 51 (61%) companies, emails can be accessed through a web login. In 28 (34%) cases, the employees do not need to know their password to access their emails, e.g., because of pre-configured mail clients.

On average, it took the participants 11 minutes to complete the survey.

4.2 General Authentication Setting

Of our 83 participants, 68 (81.93%) reported the use of company wide accounts of which all were secured at least with passwords. Ten (12%) companies additionally made use of biometric authentication (two face recognition, seven fingerprint and one palm vein recognition). One mentioned that face recognition is allowed on mobile devices. 29 (35%) participants stated that they use hardware tokens in addition to passwords. Eight (10%) participants reported they offer authentication with passwords, biometrics and tokens. Apart from this, two (2%) participants mentioned (device) certificates.

15 (18%) of the surveyed participants do not use company-wide accounts. These participants answered questions regarding their companies' email passwords. We will take a closer look at these policies in Section 4.3.2.

4.3 Password Composition Policies

In the following, all presented results only refer to PCPs used for the companies' user accounts (for regular employees), unless stated otherwise ($n = 68$).

63 (93%) of the participants stated that users are allowed to set their own account password. Two (3%) mentioned that the password is given to the user and cannot be changed by themselves. We could not find any standing out property of these two participants. In two (3%) cases, it is explicitly mentioned that an initial password is generated by the system and is changeable later; one company directly demands a change.

From the 68 participants who use company-wide accounts, we were able to extract 64 password composition policies. The remaining participants did not define a policy but gave a general description of how a policy could look. 59 (92%) of the policies get enforced technically. In two of the four companies, where this is not the case, participants mentioned that they use awareness trainings for their employees to counter the problem of common passwords.

Twenty-nine (45%) participants were part of the password policy creation process and 15 (23%) stated that the PCP was created by their predecessor.

As was expected based on the recruiting procedure, many (55%) participants who were part of the creation process of the password composition policy relied on the BSI as an influence in the PCP creation process. Figure 4 (Appendix) shows which other inspirations were used by our participants who were part of the creation process. Some other sources mentioned were ANSSI (French National Agency for the Security of Information Systems), ISO 27001, or PCI DSS (Payment Card Industry Data Security Standard). The option "Expert Panels" did not concern any specific panel but was given as a non-explicit, "consulting with experts".

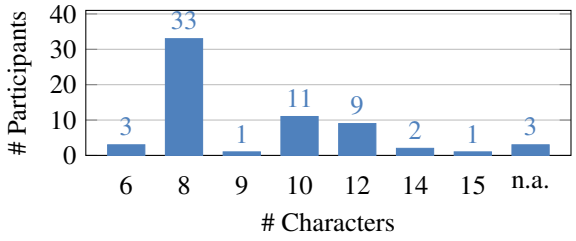


Figure 1: Minimal character length of passwords (PWA). "n.a." means no answer was given

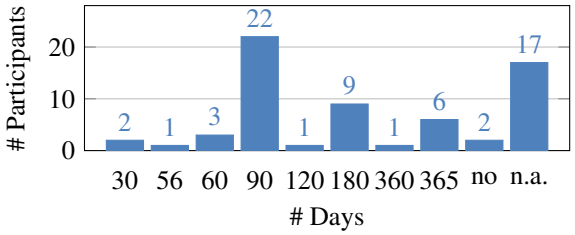


Figure 2: Password rotation cycle in days (PWA). "no" indicates participants who explicitly mentioned not using expiring passwords. "n.a." means no answer was given

4.3.1 PCP Components

In the following section, we present which elements were present in the PCPs that refer to companies' user accounts. For this, we follow the policy elements mentioned by official recommendations, as summarized in Table 2. An overview of the elements "minimal length", "password age" and "complexity" can be found in Table 4 in the Appendix.

Length Sixty-one (95%) companies use length requirements to ensure a secure password. While a minimum length is widespread (54 companies, 84%), some participants also mentioned fixed lengths (11%). However, the data for maximal and fixed length was not always clear, for example, in case of participants who stated: "Password length 8. [...]". In these cases, we count them as minimal and maximal length. In 33 (52%) companies, the participants mentioned eight characters to be their minimal password length. Eleven (17%) companies require 10 characters and 9 (14%) participants stated their minimal length to be 12 characters. Figure 1 gives an overview of how many participants mentioned which minimal length.

Password Complexity The complexity of a password depends on the number of character classes being used to create the password. For this, five different character classes can be used: uppercase letters (A-Z), lowercase letters (a-z), numbers (0-9), special characters including the space character, as well as the remaining Unicode characters that are alphabetic but not uppercase or lowercase (e.g., Chinese symbols). The latter one was only mentioned by one participant who indicated the complexity to be "Windows Password complexity", which includes all Unicode characters; so in the following, we will concentrate on the first four character classes.

Overall, 57 (89%) companies give constraints regarding the complexity. Seventeen (27%) companies require their users to build passwords using characters from all four classes. In one case, two characters of each class were demanded, one company requires a mixture between one or two characters per class, and in the other companies, one character of each class was sufficient.

Thirty-two (50%) participants mentioned that in order to fulfill their policy, characters from 3 of the four classes need to be present in a password. Fifteen (23%) specified which classes need to be covered, while 17 (27%) accepted a password as long as any three classes were present.

In five (8%) cases, participants stated that a complexity requirement is in place but did not specify, how this requirement looks.

Seven (7%) of our participants did not mention any requirements regarding complexity. However, none of them explicitly mentioned not using one.

Password Age and Password History As suggested by the BSI during the time of our study, 45 (70%) of our participants stated to force their users to change their passwords regularly. The top three rotation cycles were 90 days (34%), 180 days (14%) and 365 days (11%). Two participants explicitly mentioned not using a password expiration. While the percentage for 90 days (34%) is similar to what Habib et al. [24] found (28%), our peak at 180 and 365 days cannot be found in their sample. All password rotation cycles can be seen Figure 2.

Thirty (47%) participants reported a password history to prevent users from reusing previously used passwords. Twenty-seven (42%) of them check whether the passwords are identical, whereas three (3%) companies require significant changes, where it is, for example, not sufficient to increase a number within the old password to be accepted. Most mentioned was a history of 10 passwords (14 %) and 24 or 5 passwords (6% each). Figure 3 shows how many participants mentioned which number of previous passwords are stored.

When presented with the need to change their password and not be allowed to reuse a certain amount of their last passwords, users might counter this by changing their password several times in a row until they are allowed to use their original password again [7]. Because of this, companies use a minimal password age, as mentioned by 15 (23%) participants, so that users cannot change their password within this time period [7]. The minimal ages range from 24 hours (eleven companies) to 14 days (one company).

Allowed Characters NIST [22] recommends allowing all printing ASCII characters, the space character, and Unicode

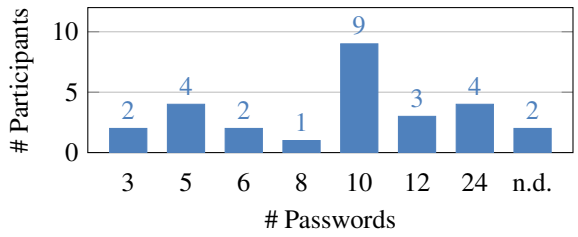


Figure 3: How many previously used passwords are not allowed to be reused (exact match). (PWA) "n.d." means no detail was given but a password history was mentioned.

characters for user-chosen passwords. As most participants only mentioned which characters are not allowed or which classes should be covered, it is hard to draw reasonable conclusions about the allowed character sets. However, one participant mentioned that a password needs to cover the "Windows Password complexity", which includes Unicode characters, e.g., "from Asian languages" [9].

Blocklists Twenty-six (41%) participants affirmed the question whether passwords were checked against common or leaked passwords. However, we did not ask for details, so we do not know how the comparison is made technically or which lists are used for this purpose.

Rarely Encountered We also found several constraints, which were only mentioned in at most three policies and were constraints to particular cases. We believe some of these are used since certain characters might break backend processing or serve as substitute for blocklists (e.g., to prevent passwords consisting of personal information such as nicknames). The atypical constraints included: (1) No colloquial language of any language, (2) No words of any language written backwards, (3) Certain special characters like € or umlauts, (4) Not more than 2 characters or sequences in series, (5) The last 20 passwords need to differ significantly from the new password, (6) Not more than 2 characters which appear in the same series in your name, (7) Not your license plate number.[2]

4.3.2 Additional Policies

In addition to PCPs, that define the user's chosen passwords on company accounts, some participants also mentioned additional PCPs, such as those for administrator accounts.

All participants who do not use company-wide accounts answered the questions regarding their email accounts.

We will present both extra sets of PCPs in the following paragraphs. Be reminded that both sets are excluded from the analysis above.

[2]While this was only mentioned by two participants, it is in fact mentioned in the implementation notes offered by the BSI [11].

General Three participants mentioned two different password composition policies. Two of them applied stricter rules if an account belonged to an administrator. In both cases, the minimal length was increased to 16 characters (while the regular accounts were required to use 12 respectively 10 characters).

One company requires its employees to use at least 20 character long passphrases for SSH and PGP/GPG keys.

Email Passwords Companies that do not use company-wide accounts for their employees were asked about their email passwords. We received 15 (18% of all responses) answers and were able to extract 13 password composition policies. Though we did not ask whether the PCPs are given by an email provider, six (46%) indicated that they were part of the creation process. The primary influence was the BSI, own knowledge, and expert panels (each one mentioned three times).

The median minimal length mentioned by participants in this group was 10 characters (range from 8 to 30 characters). Three participants mentioned very large minimal lengths: One (8%) needed 20 characters, and one (8%) only accepts passwords if they are 30 or more characters long. Additionally, the modes of the complexity were 3 and 4 character classes, and the median of the rotation cycle was 180 days (range from 180 to 365 days).

Interestingly, three of the email participants mentioned that employees generate their passwords with a password generator, whereas only one participant from the account group said so.

We also found one company that differentiated between regular and administrator accounts. They increased the minimal length from 12 to 16 characters and decreased the maximum age from 90 days for regular employees to 45 days for administrator accounts.

4.4 Effects of Authentication Methods

We asked participants how they rated password, biometric and token authentication concerning their security and usability impact and how often they encounter problems with the systems (e.g., forgotten passwords or lost hardware token). Additionally, we asked for their overall satisfaction with all used authentication methods combined. Following related work (Section 2), we assumed that unusable policies would lead to more problems and eventually to a lower satisfaction of the responsible person.

4.4.1 Influence on Security and Usability

Figure 5 (Appendix) shows what influences the use of the PCPs (n=64), biometrics (n=10), and hardware token (n=29) has on the sensed security and usability of the authentication system as well as how often problems arise for each method.

It can be seen, according to the participants, passwords lead to problems more often and form the lower bound of usability and security.

However, when comparing the scores, one has to keep in mind that biometrics and token were never used alone but always in combination with passwords. As we first asked for details about the password policies and only later for details of biometrics and token, the observed scores of biometrics and token may be compared to the security and usability of passwords. When looking at the number of problems arising from the authentication methods, tokens seem superior to biometrics and passwords. This is similar to the results of Abbott and Patil [13], who found token to have the second-highest UX rating when comparing different 2FA mechanisms.

We could identify a negative correlation between the reported impact on the security of the policy and the problems with passwords ($Spearman - \rho = -0.41294, p = 0.00938$), so the better the perceived security, the fewer perceived problems. The connection of the perceived user-friendliness and the problems is not statistically significant after correction ($Spearman - \rho = -0.3339, p = 0.05125$).

4.4.2 Satisfaction of Authentication Methods

We asked participants how satisfied they are with the present overall authentication system. Figure 6 (Appendix) shows the observed scores depending on whether a company offers passwords alone or in combination with a token. Due to small numbers, we excluded those participants using passwords, biometrics, and token (n = 8) and participants making use of passwords and biometrics (n = 2). It seems that participants using passwords in combination with tokens are more satisfied than participants using only passwords.

When concentrating on passwords, participants who stated that they created the password composition policy on their own had a median satisfaction of 4.00 (average: 3.41, sd: 0.95), while participants who were not part of the password policy creation had a median satisfaction of 3.00 (average: 2.94, sd: 0.92). This was not statistically significant (Mann–Whitney U test: $U = 624.5, p = 0.21824$).

However, a negative correlation can be reported between the number of problems with passwords and the satisfaction with the overall authentication system ($Spearman - \rho = -0.38639, p = 0.01409$); the fewer the problems, the more satisfied were the participants.

4.5 2-Factor Authentication (2FA)

If the participants stated to use more than one mechanism to authenticate accounts, we asked whether the methods are used in combination, for example, for 2-factor authentication. Of the 35 participants whose companies allow other methods than passwords, 18 stated that the methods are used as multiple factors and five companies leave the decision which method to use for authentication up to the employees.

5 Discussion

In the following section, we discuss the results. For this, we first compare the PWA PCPs to recommendations given by organizations like the BSI or NIST. This is followed by comparing the PCPs to recommendations and lessons learned from related work. After this, we analyze factors that might have influenced the PCPs when being created.

5.1 Compliance with Recommendations and Usability

5.1.1 NIST and BSI

In the following section, we discuss and compare the password composition policies and their elements with recommendations by NIST and the BSI, as seen in Table 2. We can only compare elements that are concretely covered by the corresponding guideline. The old BSI [5] recommendations do not make concrete recommendations regarding the length, complexity, or minimal password age and only require them to be “sufficient”, or “appropriate”. The same applies to the element “Quality”, which was introduced with the new BSI recommendations [6] and is expected to be “appropriate”. As the implementation notes [11] give concrete examples we compare those to the policies.

We again want to point out that the old BSI recommendation included a password expiration period and a complexity requirement during the time of the study. At the same time, NIST did not have such requirements.

Anecdotally, one participant mentioned following guidelines given by the PCI DSS (Payment Card Industry Data Security Standard) but does not consider it reasonable.

Length NIST recommends at least 8 characters for user chosen passwords. Thirty-three (52%) policies exactly fulfill this part and 57 (89%) require 8 or more characters. Three (5%) companies go against this recommendation and require only six characters.

We cannot make any statement on the minimal maximal length of 64 characters mentioned by NIST. No participant mentioned fulfilling this, but again, this does not mean that the companies only allow shorter passwords.

Complexity NIST recommends not setting complexity requirements in the form of character classes, while the old BSI recommendation suggested using a “sufficient” complexity. Fifty-seven (89%) participants mentioned complexity requirements, most often with the necessity to include three character classes in the password, so the vast majority of the companies were more in line with the old BSI recommendations than

with NIST. In the newer BSI [6] recommendation, the explicit mentioned need to establish mandatory rules of the requirement of some complexity was removed,[3] and only included that a password needs to have some level of "quality". We thus believe that our dataset can serve as a baseline for future studies observing the development of policies in companies.

Independent of what a policy requires, Tan et al. [35] found that users tend to include more character classes in their passwords.

Length and Complexity As mentioned before, the implementation notes of the BSI [11] give examples of how to combine length and complexity requirements. There are two we can use: First, a length requirement of 8 to 12 and 3 character classes. Twenty-seven (42%) participants match this. Second a length requirement of 20 to 25 and 2 character classes. None of the reported policies is equal to that example.

Password Age At the time the study was conducted, the BSI recommended regular password changes. Forty-five (70%) companies follow this recommendation and force users to change their passwords regularly. In contrast, NIST updated their recommendations in 2017, in line with results from research [39], and since advised against regular password change. Instead, a change should be forced whenever there is evidence of a compromise [22]. The new BSI recommendations are in line with this, but still recommend regular password changes, in case password cannot be checked against compromises [11].

Two (3%) participants explicitly mentioned foregoing password rotation.

Regular password changes cause users to develop mechanisms to make these changes less painful. These mechanisms, in turn, can lead to new rules added to the policies. One example of this are password histories, meaning users cannot reuse a certain number of their passwords. While most companies mentioned that the passwords as a whole are not allowed to be reused, three (4.69%) participants mentioned that significant changes are necessary. We did not ask for details of the technical implementation. The naive solution for this would be to store all passwords in plain text. This is obviously not a good idea, and systems that can be used to mitigate this issue have been proposed [17].

Another example of added rules due to a regular password change are minimal ages for passwords where users are not allowed to change their passwords within a specific period, to prevent them from cycling through and going back to their old password right away. However, this rule has the negative side-effect that users cannot change their passwords even if they assume it was compromised. Most likely, administrators could still make changes to the users' passwords instead of the users themselves, but it adds an additional step to the process. Minimal ages were mentioned by 15 (23%) participants.

Since regular password changes can cause a number of follow-up problems, it seems good that NIST and BSI no longer recommend this, and it will be interesting to see how and when companies adopt this change.

[3] Although the newer BSI [6] recommendations do not mention the complexity requirement as a mandatory rule, rules making use of complexity and length are used in the examples of the implementation notes [11].

5.1.2 Recommendations from Related Work

As summarized in Section 2, Shay et al. [34] tested 15 PCPs with over 20,000 participants. Tan et al. [35] tested 21 policies with over 6000 participants. Both research projects examined the strength of the policies by measuring the password guessability and the usability by looking at the user sentiment, the dropout rates, creation, and recall. Based on their findings, they gave recommendations for service providers at which we will have a detailed look in the following. It has to be noted that their recommendations are based on only two studies, and thus, further research is needed to explore the discovered aspects further. Additionally, not all elements that we discovered in the policies were studied or mentioned in the recommendations.

Avoid Using Length-Only Requirements Some of the tested policies by Shay et al. [34] only included length requirements. These policies seemed to be usable, and while some of the resulting passwords were quite strong, many others were very weak. Therefore, the authors suggest introducing further requirements, even if the minimal length is high. When looking at the three elements "minimal length", "complexity" and "maximum age", we only found two participants who only made statements about the length. One requested a minimal length of 6 characters but demanded the password not to include parts of the username or character/number sequences. The other company uses a minimal length of 12 characters and also forbids easy-to-guess passwords. It has to be researched what effect these additional elements have on the resulting passwords.

Do not Concentrate on Character Classes Tan et al. [35] included three policies that only contained length and complexity requirements. They found that the resulting passwords could be cracked with equal success rates, independent of the number of character classes required in the password. The authors conclude that users, at least when seeing a password meter, tend to choose longer and more complex passwords than required. When using a large blocklist, additional character class requirements do not seem to have any positive effect. The authors also found that with the use of a minimum-strength requirement, it is more usable to increase the length requirement or minimum-strength threshold compared to requiring more character classes to defend against offline attacks. In our sample, 61 of the 64 companies that also use a centrally managed user account and reported a pol-

icy mentioned using a character class requirement. Twenty-eight (46%) of them additionally mentioned to use blocklists.

If You are Using Comp8, Replace It. The PCP "comp8" included at least 8 characters, a complexity of 4, and no dictionary words. When testing this very common policy against others, Shay et al. [34] found three other PCPs to be more usable and more secure at the same time: "2class12" (minimal length of 12 characters, complexity 2), "3class12" (minimal length of 12 characters, complexity 3), and "2word16" (minimal length of 16 characters, at least 2 words). Looking at our sample, we find 10 (16%) participants who mentioned policies that match the "comp8" policy. Contrary, only four (6%) participants make use of "3class12" and no policy matches "2class12" or "2word16". It is interesting to see that while research offers good alternatives, many companies do not seem to adopt them.

Blocklists Some of the policies tested by Shay et al. [34] prohibited passwords from containing substrings from a prespecified list. They noted that this seemed to make creating a new password more difficult. However, this did not apply to the recall of passwords. Thus, they argue that including substring blocklists in the PCP is suitable if passwords do not expire too often. Tan et al. [35] found differences between different wordlists and matching algorithms (e.g. case-insensitive). They recommend not combining blocklist and character class requirements, especially when any password is rejected that exactly matches one included in a public leak.

Of the participating companies, 32 (50%) either confirmed whether they check the user-chosen passwords against commonly used passwords or mentioned to prohibit users from using certain sequences in their passwords as, for example, the company name, character/numerical sequences or dictionary words. Twenty-five (78%) of them additionally require regular password changes, that Shay et al. consider as an unfavorable combination [34].

Twenty-eight (46%) companies used a blocklist in combination with complexity requirement, which is not recommended by Tan et al. [35]. We do not have further insight into (a) what lists were used nor (b) how the comparison takes place (are numbers and digits included in a full string comparison / is the comparison case-insensitive?). Since different implementations seem to affect the security of the resulting passwords, but also on the time needed to create passwords, as found by Tan et al. [35], we believe future work should look at the actual implementation of blocklists within companies.

Minimum-Strength Requirements According to Tan et al. [35], minimum-strength requirements (i.e., number of guesses needed) are beneficial for password creation and, at the same time, result in strong and easy to remember passwords overall. When needing protection against offline attacks, the authors recommend their so-called "1c12+NN10" policy. In comparison to blocklist requirements, minimum-strength policies seem to combine better protection with improved usability.

None of our participants explicitly mentioned checking the passwords against a guessing attack. 5 participants required not to use "easy to guess" passwords. However, they did not specify if and how this is checked. It could thus be that the user is not supported fulfilling this rule. It remains to be seen if and how this relatively new knowledge about minimum-strength requirements will be applied in PCPs within the following years.

Ur et al. [37] recommended supporting users in developing good approaches for creating passwords and teaching them to correctly judge their decisions regarding password strength instead of creating PCPs that focus on character-class structures. Two of the participants explicitly mentioned awareness trainings for their employees regarding passwords. We did not ask about this explicitly, so there are probably more participants who, in fact, use awareness trainings.

Inglesant et al. [25] studied the effect of the PCP of two different companies, from which one had a very complex policy. It required their employees to use passwords consisting of 7 to 8 characters, a complexity of 3, no dictionary words, not exchanging o with 0 or i with 1, an expiration of 120 days, and significant changes to the 12 previous passwords. Many participants from this company expressed frustration and negative feelings towards password creation and recall. The authors note that the unusability arises from the combination of complexity, regular password changes, and the necessity to make significant changes to a previous password. Even though they conducted the study 10 years ago, we still saw many policies that have the potential to create user frustration as they consist of many elements that all have to be kept in mind when creating the password (cf. section 4.3.1). We hope that the revised BSI recommendations help in creating more user-friendly password composition policies.

5.2 Factors

Before conducting the survey, we had two themes that we assumed would influence the PCPs in companies:(1) **company size**: Tiefenau et al. [36] showed a significant difference between small and large organizations regarding formal update processes. We thus hypothesized that this difference could also be seen in other areas. (2) **the consulted recommendations (e.g., BSI or NIST)**, as they differ in details (cf. Section 2.3).

5.2.1 Size of Company

Table 3 shows that the amount of clients managed by the participant grows with the company size based on the number of employees. The amount of participants reporting a company-wide account does not increase with the number of employees.

Although the percentages at the two poles are clearly different, the small number of companies per bin does not allow us to draw conclusions. But we find that the use of a company-wide account does not seem like something extraordinary, even for small companies.

When separating the small and medium-sized companies (G_small, n=23) from the big companies (G_big, n=40) according to the definition from the EU [3], we do not see a big difference in the PCPs. They do not vary much in terms of the mean length (G_small = 9.8 vs. G_big = 8.95) or the mean maximum age (G_small = 156.8 vs G_big = 141). We also found no difference in the modes of the complexity when simplified to only the number of required character classes (1,2,3,4): G_small = 3 vs. G_big = 3.

5.2.2 Consulted Recommendations

As already touched in Section 4.3 and can be seen in detail in Figure 4, the participants reported using different sources as a basis for their policies. In 19 out of 29 cases, the participants listed more than one source of information besides 'Own Knowledge', statements from the 'Other' text field included. 2 participants reported only one institution besides 'Own Knowledge'.

Slightly more than half of the participants who took part in the creation of the PCP (16, 55%) used the BSI as a basis for the policy. Fourteen of those were recruited via the BSI newsletter, so that we might see a recruitment bias here.

When looking at the minimal length, we could not identify a big difference between those who use the BSI as an inspiration and were part of the creation process (G_BSI, n = 16) and those who do not (G_nBSI, n = 13): We found a mean minimal length of 10.5 characters for G_BSI and of 9.7 characters for G_nBSI. Nevertheless, we saw deviating means for the maximum age of passwords (225.00 days for G_BSI, 135 days for G_nBSI). The BSI guidelines suggested a regular password change. The high number could be based on the will to mitigate this measure but does not explain why the mean number of days for G_nBSI is so low.

Potentially, this points to the need to further separate the participants into groups classified by aspects like industrial sector or based on a risk evaluation.

The mode of the complexity is 3 for both groups.

5.2.3 Self-Made Policies

During analysis of the data, we found that a higher amount of policies that require a minimal length of 8 characters were mentioned by participants who were not part of the creation process. (31% in selfmade vs. 69% in not-selfmade). Overall, the mean minimal length of the not-self-made PCPs is also slightly lower (10.1 characters vs. 8.5 characters). A possible explanation might be the time a policy was created and official recommendations during those times.

Most (24 of 29) of the participants who helped create the policy reported having based the policy on 'Own Knowledge'. Future research should investigate this further as it is open to discuss whether this is a situation of "writing your own crypt library" or not. It should be noted that the used Software often provides options for policies, and so this probably heavily influences the policies (e.g., through default values).

PCPs of participants who were part of the creation process contained more elements when compared to the PCPs of participants who stated that they were not involved. This mainly included forbidden password elements as not using the username or license plate numbers. This may be an artifact of the methodology (e.g., recall bias).

While it is tempting to see a causal relationship here, be aware that the hypotheses around the self-made aspect are build from the data, so this phenomenon should be investigated in a separate study.

5.3 Heterogeneity

One of the main observation we made is that there is large heterogeneity in the landscape of PCPs, that we reported in Section 4.3. All PCPs used for company-wide accounts arrange in the area of 6-15 characters minimum (with a clear peak at a minimal character length of 8, Figure 1), an expiration range from 30 to 365 days (with two peaks at 90 and 180 days, Figure 2) and a password history of 3 to 24 passwords (with a peak at a history of 10, Figure 3). Yet we only found two PCPs that were identical concerning all mentioned elements as password history, forbidden words, etc. As this could be an artifact of our methodology, as discussed in section 6, we focused on the most common combinations of the three elements "complexity requirement", "minimal length" and "password rotation" as well as the number of PCPs that mentioned this combination (cf. Table 1). As mentioned in Section 4.3.1, some participants specified which character classes need to be covered in case they had the complexity requirement "3 out of 4". We merged all of them for Table 1 and only looked for the number of required character classes. Using the three elements, we found 41 out of 540 (9 different minimal lengths * 10 different maximum ages * 6 different numbers of required character classes (1 to 4, not stated, unspecific) possible combinations in the PWA policies

6 Limitations

Our study, like most surveys, has several important limitations.

The participants were invited over the newsletter sent by the BSI. Therefore, most of our participants are likely to be already interested in security-relevant topics. As the participation was voluntary and not remunerated, this effect is even heightened. As already stated in Section 3.3, we do not know for sure whether the participants were responsible for the

Complexity At least one char	Min. Length	Max. Age	Policies
3 classes	8	90 days	7
4 classes	8	90 days	5
3 classes	10	90 days	4
4 classes	8	n.a.	3
3 classes	8	180	3
4 classes	12	n.a.	2
4 classes	10	90	2
3 classes	8	60	2
3 classes	8	365	2
3 classes	8	n.a.	2
3 classes	12	365	2
		Sum:	34 of 64

Table 1: Most common PCP element combinations of complexity (at least one character of each class), minimal length and maximal age.
Policies gives the number of policies that showed this element combination. All other combinations only occurred once in the data set.

PCPs. It is also possible that multiple participants come from the same company. We searched the data for evidence but could not identify such.

As with any survey, participants may have selected the first answer that seemed appropriate without deeply thinking about their true beliefs and behavior. We tried to mitigate this the answers were randomized wherever meaningful. We also designed the survey to an expected time of 10 minutes.

The self-report and the recall bias has most likely affected our results. It is possible that participants answered in an effort to be more socially desirable. To reduce this bias, we kept the surveys anonymous and asked for as little demographic data as possible. It is also reasonable that they did not answer the free text answers in full detail, not because of bad faith but because they forgot it, misremembered parts, or because a detailed answer would have taken too much time to answer. This might have influenced the heterogeneity within the PCPs. To compensate for this in our analysis, we thus separated not mentioning something and explicitly excluding something. We believe this leads to a more optimistic estimation of the policies.

Due to differences in company structures, there are likely questions that are not in the direct area of responsibility of the participants among the diverse set of questions asked. We tried to mitigate this with a clear statement at the start of the survey to ensure that at least password policies are present in this area. If the participant did not know the answer, there was the answer option “I do not know”.

As a consequence of the heterogeneity of the PCPs, it is difficult to compare them. Many of our comparisons happen on large groups, only taking a few possible elements into account, disregarding their combination.

7 Conclusion

We conducted a survey to evaluate the current status quo of password composition policies in German companies. Our main finding is that there is high heterogeneity in the PCPs. While we cannot draw causal conclusions, it seems likely that the clashing nature of the national (BSI) and international guidelines (NIST) and the vague nature of the national BSI guidelines contribute to this heterogeneity. There is also a high prevalence of PCP elements that the research community, as well as NIST consider harmful, such as password expiry and character class requirements. When comparing the PCPs to recommendations made by related work, we found many that are very likely to be user-unfriendly. While the new BSI guidelines might fix the latter issue, they still require an analysis of the companies situation and leaf room for interpretation. Thus we do not believe that they will reduce the heterogeneity. While heterogeneity itself is not necessarily harmful and could even have security advantages, we doubt that the policy differences are based on conscious decisions and believe that they are more likely the product of a best-effort process. We recommend further research in this area and test whether more concrete recommendations lessen the burden on decision-makers within companies, who currently have to make many decisions that require a high level of domain knowledge.

8 Future Work

As mentioned in Section 2.3, the BSI changed their recommendation after the first survey was conducted. In future work, we will monitor how this guideline change affects the PCPs of companies.

Most of our findings indicate trends that need further research with other methodologies to validate them. We encountered several PCP elements that can not easily be enforced technically. We plan to examine whether custom enforcement mechanisms were implemented or whether the hope is that users follow the instructions even though they are not enforced. It is also open how users perceive the difference. While most policies were found to be similar in the big picture, they differed in their details. While this may be for good reasons, it also shows how diverse the landscape is, and further research is needed to see whether this is needed or if the benefit of one usable policy is higher.

The surveyed participants were all employees of German companies. Further research has to be conducted to validate the findings across cultures and study the influence of different local recommendations.

We were able to show a difference in the PCPs reported by their creators compared to PCPs created by somebody else than the participant. Still to be researched is the process of creating this policy, how big the personal factor is (and should be), and what role official recommendations play.

Acknowledgments

This work was partially funded by the Werner Siemens Foundation.
We thank Christian Tiefenau, Dirk Backofen, and Alexander Häring for their help, domain knowledge, and input. We thank our reviewers and shepherd, who helped improve the paper a lot, and all participants who took the time to answer our questions.

References

[1] About | Bitkom e.V. https://www.bitkom.org/EN/About-us/About-us.html. Accessed: May 31, 2021.

[2] Authentication cheat sheet. https://cheatsheetseries.owasp.org/cheatsheets/Authentication_Cheat_Sheet.html. Accessed: May 31, 2021.

[3] COMMISSION RECOMMENDATION of 6 May 2003 concerning the definition of micro, small and medium-sized enterprises. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32003H0361. Accessed: May 31, 2021.

[4] Cyber Security Cluster Bonn. https://cyber-security-cluster.eu/. Accessed: May 31, 2021.

[5] IT-Grundschutz Compendium - Final Draft, 1 February 2019. https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/Grundschutz/International/bsi-it-gs-comp-2019.pdf?__blob=publicationFile&v=1. Accessed: May 31, 2021.

[6] IT-Grundschutz-Kompendium Februar 2020. https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Grundschutz/Kompendium/IT_Grundschutz_Kompendium_Edition2020.pdf?__blob=publicationFile&v=1. Accessed: May 31, 2021.

[7] Minimum password age. https://docs.microsoft.com/en-us/windows/security/threat-protection/security-policy-settings/minimum-password-age. Accessed: February 27, 2020.

[8] Nur fünf Zeichen fürs Banking-Passwort? https://www.heise.de/-4935773. Accessed: June 02, 2021.

[9] Password must meet complexity requirements. https://docs.microsoft.com/en-us/windows/security/threat-protection/security-policy-settings/password-must-meet-complexity-requirements. Accessed: May 31, 2021.

[10] Qualtrics. https://www.qualtrics.com. Accessed: May 31, 2021.

[11] Umsetzungshinweise zum Baustein: ORP.4. Identitäts- und Berechtigungsmanagement. https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Grundschutz/Umsetzungshinweise/Umsetzungshinweise_2021/Umsetzungshinweis_zum_Baustein_ORP_4_Identitaets_und_Berechtigungsmanagement.pdf?__blob=publicationFile&v=1. Accessed: May 31, 2021.

[12] Zuordnungstabelle ISO zum modernisierten IT-Grundschutz. https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Grundschutz/Kompendium/Zuordnung_ISO_und_modernisierter_IT_Grundschutz.html?__blob=publicationFile&v=1. Accessed: May 31, 2021.

[13] Jacob Abbott and Sameer Patil. How mandatory second factor affects the authentication user experience. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–13, New York, NY, USA, 2020. Association for Computing Machinery.

[14] Jeremiah Blocki, Saranga Komanduri, Ariel Procaccia, and Or Sheffet. Optimizing password composition policies. In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, EC '13, page 105–122, New York, NY, USA, 2013. Association for Computing Machinery.

[15] Joseph Bonneau and Sören Preibusch. The Password Thicket: Technical and Market Failures in Human Authentication on the Web. In *WEIS*, 2010.

[16] John L. Campbell, Charles Quincy, Jordan Osserman, and Ove K. Pedersen. Coding in-depth semistructured interviews: Problems of unitization and intercoder reliability and agreement. *Sociological Methods & Research*, 42(3):294–320, 2013.

[17] Rahul Chatterjee, Joanne Woodage, Yuval Pnueli, Anusha Chowdhury, and Thomas Ristenpart. The typtop system: Personalized typo-tolerant password checking. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pages 329–346, New York, NY, USA, 2017. Association for Computing Machinery.

[18] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[19] Juliet M. Corbin and Anselm Strauss. Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology*, 13(1):3–21, 1990.

[20] Joseph L. Fleiss, Bruce Levin, and Myunghee C. Paik. *Statistical methods for rates and proportions*. john wiley & sons, 2013.

[21] Dinei Florêncio and Cormac Herley. Where do security policies come from? In *Proceedings of the Sixth Symposium on Usable Privacy and Security (SOUPS)*, New York, NY, USA, 2010. Association for Computing Machinery.

[22] Paul A. Grassi, James L. Fenton, Elaine M. Newton, Ray A. Perlner, Andrew R. Regenscheid, William E. Burr, Justin P. Richer, Naomi B. Lefkovitz, Jamie M. Danker, Yee-Yin Choong, Kristen K. Greene, and Mary F. Theofanos. NIST Special Publication 800-63b: Digital Identity Guidelines. https://doi.org/10.6028/NIST.SP.800-63b. Accessed: May 25, 2021.

[23] Hana Habib, Jessica Colnago, William Melicher, Blase Ur, Sean Segreti, Lujo Bauer, Nicolas Christin, and Lorrie Cranor. Password creation in the presence of blacklists. *Proc. USEC*, page 50, 2017.

[24] Hana Habib, Pardis E. Naeini, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Nicolas Christin, and Lorrie F. Cranor. User behaviors and attitudes under password expiration policies. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS) 2018*, pages 13–30, Baltimore, MD, August 2018. USENIX Association.

[25] Philip G. Inglesant and Martina A. Sasse. The true cost of unusable password policies: password use in the wild. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 383–392, New York, NY, USA, 2010. Association for Computing Machinery.

[26] Patrick G. Kelley, Saranga Komanduri, Michelle L. Mazurek, Richard Shay, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie F. Cranor, and Julio Lopez. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In *2012 IEEE Symposium on Security and Privacy*, pages 523–537. IEEE, 2012.

[27] Saranga Komanduri, Richard Shay, Patrick G. Kelley, Michelle L. Mazurek, Lujo Bauer, Nicolas Christin, Lorrie F. Cranor, and Serge Egelman. Of passwords and people: Measuring the effect of password-composition policies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2595–2604, New York, NY, USA, 2011. Association for Computing Machinery.

[28] Peter Mayer, Jan Kirchner, and Melanie Volkamer. A second look at password composition policies in the wild: Comparing samples from 2010 and 2016. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS) 2017*, pages 13–28, Santa Clara, CA, July 2017. USENIX Association.

[29] Michelle L. Mazurek, Saranga Komanduri, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie F. Cranor, Patrick G. Kelley, Richard Shay, and Blase Ur. Measuring Password Guessability for an Entire University. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, CCS '13, pages 173–186, New York, NY, USA, 2013. Association for Computing Machinery.

[30] Sören Preibusch and Joseph Bonneau. The password game: Negative externalities from weak password practices. In *International Conference on Decision and Game Theory for Security*, pages 192–207. Springer, 11 2010.

[31] Tobias Seitz, Manuel Hartmann, Jakob Pfab, and Samuel Souque. Do Differences in Password Policies Prevent Password Reuse? In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI '17, pages 2056–2063, 05 2017.

[32] Richard Shay and Elisa Bertino. A comprehensive simulation tool for the analysis of password policies. *International Journal of Information Security*, 8(4):275–289, 08 2009.

[33] Richard Shay, Saranga Komanduri, Adam L. Durity, Phillip Huh, Michelle L. Mazurek, Sean M. Segreti, Blase Ur, Lujo Bauer, Nicolas Christin, and Lorrie F. Cranor. Can long passwords be secure and usable? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 2927–2936, New York, NY, USA, 2014. Association for Computing Machinery.

[34] Richard Shay, Saranga Komanduri, Adam L. Durity, Phillip S. Huh, Michelle L. Mazurek, Sean M. Segreti, Blase Ur, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Designing password policies for strength and usability. *ACM Transactions on Information and System Security (TISSEC)*, 18(4):13, May 2016.

[35] Joshua Tan, Lujo Bauer, Nicolas Christin, and Lorrie F. Cranor. Practical recommendations for stronger,

more usable passwords combining minimum-strength, minimum-length, and blocklist requirements. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, page 1407–1426, New York, NY, USA, 2020. Association for Computing Machinery.

[36] Christian Tiefenau, Maximilian Häring, Katharina Krombholz, and Emanuel von Zezschwitz. Security, availability, and multiple information sources: Exploring update behavior of system administrators. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pages 239–258. USENIX Association, August 2020.

[37] Blase Ur, Fumiko Noma, Jonathan Bees, Sean M. Segreti, Richard Shay, Lujo Bauer, Nicolas Christin, and Lorrie F. Cranor. "I Added'!'at the End to Make It Secure": Observing Password Creation in the Lab. In *Eleventh Symposium On Usable Privacy and Security (SOUPS) 2015*, pages 123–140, Ottawa, July 2015. USENIX Association.

[38] Ding Wang and Ping Wang. The emperor's new password creation policies. In *European Symposium on Research in Computer Security*, pages 456–477. Springer, 11 2015.

[39] Yinqian Zhang, Fabian Monrose, and Michael K. Reiter. The security of modern password expiration: An algorithmic framework and empirical analysis. In *Proceedings of the 17th ACM conference on Computer and communications security*, CCS '10, pages 176–186, New York, NY, USA, 2010. Association for Computing Machinery.

Appendix

A Survey

Consent

Welcome!

Thank you for taking time to participate in our study. The study is conducted by the team of Prof. Dr. Matthew Smith at the University of Bonn.

In this Study we want to find out more about the current state of authentication methods, in particular password policies in various companies. We will not ask for any personal information or data that could identify your company. Further, we will only report anonymous aggregated information. The goal of our research is to identify the needs of industry and develop supporting measures to increase IT-security. With your participation you will make an valuable contribution to this goal.

The survey is addressed to persons who are responsible for authentication and password policies in companies.

The survey will take 5-10 minutes, is voluntary and can be canceled at any time. If you have any questions, please contact -mail-.

By continuing with the study, you confirm that you are at least 18 years old and consent to your data being used anonymously. As the data is collected anonymously, it is not possible to delete any data after taking the survey.

Accounts

- Is there a company-wide account per user, that is managed centrally? (E.g., for logging into the workstation, communication platform, email or the like.)
  Yes / No
  - If yes:
    * What can this account be used for? (Multiple answers possible.)
      Email / Workstations / Communication platform (SharePoint, Slack etc.) / VPN into corporate network / Access to shared corporate data (e.g., Active Directory) / Other: [Free text]
    * Which methods can be used to log in? Please check the applicable.
      Password or PIN / Biometrics (e.g., Fingerprint, Face recognition) / Hardware Token (e.g. Smartcard, Token, Smartphone)
    * In there any other method in use that it not listed?
      Yes, the following: [Free text]/ No
    * You stated, that there are several methods in use which enable your employees to log in. Are the methods used in combination (e.g., 2FA)?
      Yes, the methods are used in combination (2FA) / No, the employees can choose one of the methods / Other: [Free text]/ I do not know / I do not wish to make a statement

Passwords

You stated that there is no company-wide account with wich the employees can log into several services.
The following questions regard the email accounts of your employees and their passwords (Webmail, imap, pop, etc.).
Or
You stated, that your employees use passwords/PINs to log in. The following questions regard these passwords/PINs.

- How are passwords handled?
  Users can choose them themselves / Passwords are created by a system, and users cannot change them / I don't want to make a statement / Other: [Free text]

- What specification (also called password policy) do passwords need to fulfill (e.g., at least x characters, new password needs to be selected after x days, etc.)

  This question is the main focus of our research. Please be as detailed as possible. If possible and allowed, please copy your specification into the following text box. At this point we want to remind you, that the data is managed anonymously. It will not be possible to identify your company.
  [Free text]

- Are these specifications enforced by the system?
  Yes / No / There are no specifications / I do not know / I do not wish to make a statement / Partially:[Free text]

- Optional: What reasons spoke against the introduction of a password policy?
  [Free text]

- Are users prevented from picking passwords that belong to the most common passwords?
  Yes / No / Other:[Free text] / I do not know / I do not wish to make a statement

- Who created the specifications (password policies) for the passwords?
  Myself / My predecessor / Somebody else: [Free text]/ I do not know / I do not wish to make a statement / There are no specifications

- What are the specifications based on? (Multiple answers possible.)
  Own knowledge / Expert panels / Exchange with other companies / NIST (National Institute of Standards and Technology) / BSI (Bundesamt für Sicherheit in der Informationstechnik) / OWASP (Open Web Application Security Project) / Other: [Free text]/ I do not know / I do not wish to make a statement

- How do the password policies impact the user-friendliness of the authentication system?
  1: Very negative – 5: Very positive

- How do the password policies impact the security of the authentication system?
  1: Very negative – 5: Very positive

- How often do passwords cause problems in your company (e.g., forgotten passwords, etc.)?
  1: Very rarely – 5: Very often

- Is there a policy which specifies how the passwords are stored in the system (hash function, length of the salt, etc)?
  Yes / No / I do not know / I do not wish to make a statement

- Is there a process which initiates an update of the policy on how to store passwords?
  Yes / No / I do not know / I do not wish to make a statement

- Optional: How are stored passwords protected? We are particularly interested in the hash and salt functions which are used.

  We want to remind you that the data is gathered anonymously and we are not able to link it to your company.
  [Free text]

Biometrics

You stated, that your employees use biometrics to log in. The following questions regard this method.

- What kind of biometrics are in use?
  Fingerprint / Iris / Face recognition / Other: [Free text]/ I do not wish to make a statement

- How does the biometric authentication impact the user-friendliness of the authentication system?
  1: Very negative – 5: Very positive

- How does the biometric authentication impact the security of the authentication system?
  1: Very negative – 5: Very positive

- How often does the use of biometric authentication cause problems?
  1: Very rarely – 5: Very often

- Optional: Do you wish to provide us with additional information about this topic?
  [Free text]

Hardware Token

You stated, that your employees use a hardware token to authenticate. The following questions regard this token.

- Does the token support FIDO2?
  Yes / No / I am not sure / I do not wish to make a statement

- How does the token impact the user-friendliness of the authentication system?
  1: Very negative – 5: Very positive

- How does the token impact the security of the authentication system?
  1: Very negative – 5: Very positive
- How often does the usage of the token cause problems?
  1: Very rarely – 5: Very often
- Optional: Do you wish to provide us with additional information about this topic?
  [Free text]

Demographics

- Please check the conditions which apply to your company. (Multiple answers possible.)
  There are employees who can access their emails outside the company network /
  There are employees who can access their emails using a weblogin /
  There are employees who do not need to know the password for accessing their emails, e.g., as the email-client is pre-configured
- Is there any additional security for emails? (e.g., encryption in combination with a smartcard)
  Yes, obligatory / Yes, voluntary / No / I do not wish to make a statement
- How many employees work in your company?
  1-9 / 10-49 / 50-249 / 250-499 / 500-999 / 1000 or more / Not sure / I do not wish to make a statement
- How many desktop clients do you manage?
  1-9 / 10-49 / 50-249 / 250-499 / 500-999 / 1000 or more / Not sure / I do not wish to make a statement
- How many employees in your company work full time on IT-security topics?
  0 / 1 / 2-5 / 6-10 / 11-20 / 21 or more / Not sure / I do not wish to make a statement
- How satisfied are you with your authentication system?
  1: ☹ – 5: ☺
- Has this questionnaire motivated you to update parts of your authentication system in the near future? If yes, which parts?
  Password Policies / Security measures for stored passwords / Adding biometrics / Adding hardware token / No / Other: [Free text]

B Additional Figures and Tables

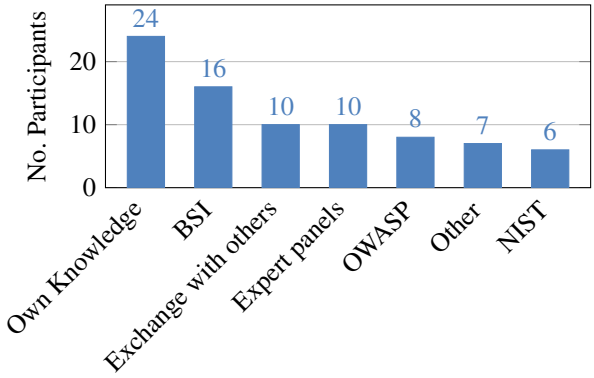


Figure 4: Sources of inspirations for the password composition policies reported from participants who took part in the creation (n=29, PWA). Multiple answers were possible.

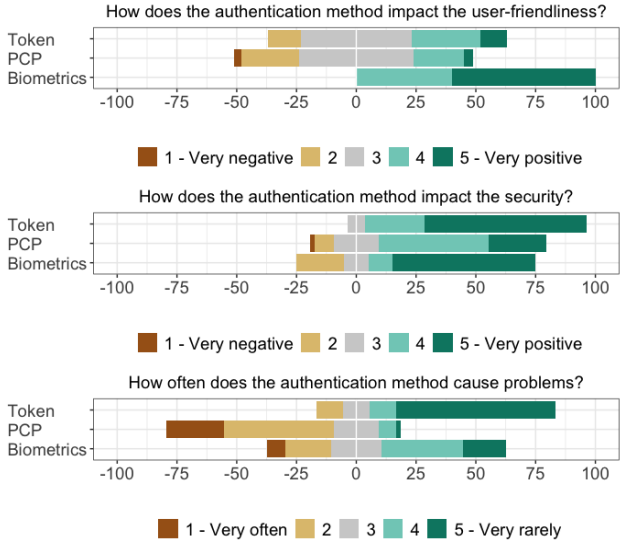


Figure 5: Impact of the different authentication methods: token ($n = 29$), the PCP ($n = 64$) and biometrics ($n = 10$) on the perceived user-friendliness, security and frequency of problems. Only PWA policies are used for the figures.

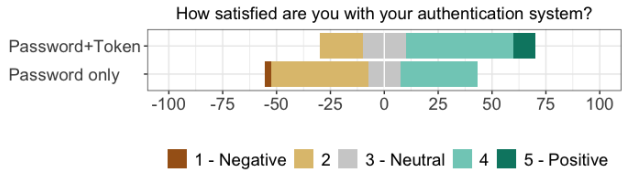


Figure 6: Satisfaction of participants with their overall authentication system, depending on methods in use: Passwords only (n = 33) and passwords in combination with token (n = 20). Numbers on the x-axis are percentages. Only PWA policies are used for this figure.

Policy Elements	**NIST** (2020) [22]	**BSI Old** (2019) [5]	**BSI New** (2020) [6]
Quality	-	-	Appropriate
Minimal length	8	Sufficient	-
Minimal maximal length	64	-	-
Complexity	Advised against	Sufficient	-
Maximal age	Advised against	Appropriate	-
Allowed characters	All ASCII & Unicode characters	-	-
Blocklist	At least: - Leaked passwords - Dictionary words - Repetitive or sequential characters - Context-specific words	-	- Easy to guess - Common passwords - Reused passwords

Table 2: Recommendations for password policies by different organizations, split by their elements. The BSI revised their recommendation in 2020.

		No. of Participants		**No. of Participants per No. of Managed Clients**						
		Company-wide	**Email**	**0-9**	**10-49**	**50-249**	**250-499**	**500-999**	**>= 1000**	**n.a.**
Company size	1- 9	5	3	6	1					
	10 - 49	7	5	2	9					1
	50 - 249	12	1			11	1			1
	250 - 499	7	2			1	7			1
	500 - 999	6	1					6	1	
	>= 1000	30	3	2				5	23	3
	No answer	1	0							

Table 3: Number of participants depending on the Company Size and Number of Managed Clients. Empty fields indicate 0. n.a. = No answer

	Element	Account Policies n= 64	Mail Policies n= 13	Additional Policies n = 4
Minimal Length	6	3	2	-
	8	33	4	-
	9	1	-	-
	10	11	2	-
	12	9	1	-
	14	2	1	-
	15	1	-	-
	16	-	-	3 (Admin)
	20	-	1	1 (Passphrases)
	30	-	1	-
	Unspecific	1	-	-
	N.A.	3	1	-
	Any minimal length	61	12	4
Maximal Age (Days)	30	2	-	-
	42	-	1	-
	45	-	-	1 (Admin)
	56	1	-	-
	60	3	-	-
	90	22	4	-
	120	1	-	-
	180	9	2	-
	360	1	-	-
	365	6	1	-
	Explicitly not	2	-	-
	N.A.	17	5	3
	Any maximal age	45	8	1
Character Classes (>= 1 each)	Special character	1	-	-
	2 (Letter, Digit)	-	1	-
	2 (Digit, Special)	2	-	-
	3 (Capital, Digit, Special)	3	-	-
	3 (Capital, Lowercase, Special)	3	-	-
	3 (Capital, Lowercase, Digit)	7	1	1 (Admin)
	3 (Letter, Digit, Special)	2	-	-
	Any 3 out of 4	16	2	-
	Any 3 out of 5 (incl. Unicode)	1	-	-
	4 (Capital, Lowercase, Digit, Special)	15	3	1 (Admin)
	4 (at least 2 each)	1	1	-
	4 (Capital and Lowercase: at least 1; Digit+Special: at least 2)	1	-	-
	Unspecific	5	1	1 (Admin)
	N.A.	7	4	1
	Any character class requirement	57	9	3

Table 4: Number of policies with the different elements of “minimal length”, “maximal age” and “character classes”. The four standard character classes are: Lowercase, Capital, Digit, and Special character.

Using a Blocklist to Improve the Security of User Selection of Android Patterns

Collins W. Munyendo*, Miles Grant*, Philipp Markert‡, Timothy J. Forman§, and Adam J. Aviv*
*The George Washington University, ‡Ruhr University Bochum, §United States Navy

Abstract

Android patterns remain a popular method for unlocking smartphones, despite evidence suggesting that many users choose easily guessable patterns. In this paper, we explore the usage of blocklists to improve the security of user-chosen patterns by disallowing common patterns, a feature currently unavailable on Android but used by Apple during PIN selection. In a user study run on participants' smartphones ($n = 1006$), we tested 5 different blocklist sizes and compared them to a control treatment. We find that even the smallest blocklist (12 patterns) had benefits, reducing a simulated attacker's success rate after 30 guesses from 24 % to 20 %. The largest blocklist (581 patterns) reduced the percentage of correctly guessed patterns after 30 attempts down to only 2 %. In terms of usability, blocklists had limited negative impact on short-term recall rates and entry times, with reported SUS values indicating reasonable usability when selecting patterns in the presence of a blocklist. Based on our simulated attacker performance results for different blocklist sizes, we recommend blocking 100 patterns for a good balance between usability and security.

1 Introduction

Restricting access to smartphones is critical for security, as these devices play an important role in our daily lives. A common method to secure smartphone access is unlock authentication, such as using a PIN or password, that the user enters to unlock the device. On Android devices, users can also choose to select a graphical method in the form of unlock patterns, where users enter a previously selected pattern by swiping on a 3x3 grid of points.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

While user-selected passwords and PINs for mobile authentication have been shown to be more resilient to guessing attacks compared to Android patterns [20], patterns remain popular among a large group of Android users. Our study finds that about 27 % of participants use patterns, matching inquiries from prior work [4, 14, 17, 20]. Even though 3x3 patterns allow for 389,112 options, more than the 10,000 choices offered by 4-digit PINs, users select from a much smaller subset of patterns that are easily predicted and would be guessed by an informed attacker, even after just 30 attempts [4, 29].

There have been several proposals in the past to improve the security of Android patterns. Some suggest using ad-hoc strength meters [2, 25, 26], feedback during selection [12], rearrangement of the contact points [28, 29], or expansion from a 3x3 to a 4x4 grid [4]. However, these suggestions all have their drawbacks. For instance, increasing the grid size has proven not to increase security significantly for an online attacker with a few guesses [4]. Other proposals, including the rearrangement of the grid, change the simple interface that makes Android patterns so popular in the first place.

To address these challenges, we propose using blocklists during pattern selection. This feature, which is used by Apple iOS devices during PIN selection, disallows common options so that users select more diversely. Recent research on mobile authentication PINs [20] and Knock Codes [22] indicates that a well sized blocklist can have significant improvements on security with limited impact on usability. In this paper, we ask *(a) what is the security and usability impact of blocklists on Android unlock patterns, and (b) what is the "right" sized blocklist that balances security and usability?*

To answer these questions, we carried out an online survey on Amazon Mechanical Turk with $n = 1006$ participants. Participants were assigned to 1 of 6 treatments, a control as well as 5 blocklist-enforcing treatments. For the latter, we varied the size of the blocklists with patterns chosen based on prior studies [4, 19, 29, 31]. In the course of the survey, participants created and recalled a pattern, answered questions about the general usability and described their strategies for selecting their pattern. Participants that encountered a blocklist were

additionally asked about changes to their selection strategy while those that did not were asked if their strategy would change upon encountering a blocklist warning.

We evaluated the security of unlock patterns selected across different treatments using guessability metrics. We primarily considered a throttled attacker scenario as it is the most relevant for mobile authentication where an attacker only has 10–30 guesses before a lockout or at least significant delays ($>$ 1 hr) occur on the device. We find that that 24 % of patterns in the control treatment are guessed after 30 guesses. In contrast, the smallest blocklist reduces the attacker's performance to 20 %, and with the largest blocklist in place, the attacker only guesses 2 % of the patterns within 30 attempts.

For usability, blocklists had minimal impact on short-term recall rates. While the average selection time increases due to the interaction with the blocklist (a one time cost), changes in entry times are negligible. Participants in the largest blocklist treatments only took, on average, an additional 0.26 seconds to enter their patterns. Participant responses evaluated using the System Usability Scale (SUS) support these findings, with scores ranging from 78.6 for the control to 71.6 for the largest blocklist treatment, indicating that the addition of blocklists improves security while appearing not to have meaningful effects on the usability of unlock patterns. However, additional work is required to explore long-term recall rates.

To summarize, we make the following contributions:

1. We study the effects of blocklists on Android unlock patterns, showing that they are able to significantly increase security even for small blocklist sizes. Patterns selected in the blocklist treatments are harder to guess for both a simulated and a perfect knowledge attacker.
2. We show that blocklists might not have meaningful effects on the usability of unlock patterns. The SUS scores, entry times, and short-term recall rates across all 5 blocklist treatments are comparable to the control treatment.
3. We provide guidance to improve the existing implementation of unlock patterns, with our results suggesting that a blocklist containing the 100 most common patterns improves the security of user-chosen patterns while appearing to minimally impact their short-term recall rates.

2 Related Work

Android unlock patterns, first introduced in 2008 as a modification of the Pass-Go scheme [27], are one of the most widely used knowledge-based authentication mechanisms on smartphones today. Despite being less secure than PINs [1, 4, 10, 20, 29, 32] or passwords [9, 23], 27 % of participants in our study use patterns to secure their smartphones, which matches inquiries from prior work [4, 14, 17, 20].

Some of the security limitations of patterns were first demonstrated by Uellenbeck et al. [29]. Through a large scale user study measuring users' actual choices of patterns, Uellenbeck et al. found that selection strategies for patterns are biased, including a preference to start from the top left corner and end in the bottom right corner of the grid. Loge et al. [19] found that personal traits of a user influence the strength of unlock patterns they select. Other studies have shown patterns to be vulnerable to smudge attacks [6, 11], shoulder surfing attacks [5, 8, 23], sensor attacks [7], video attacks [33], and physical attacks [3].

As a workaround, there have been several proposals to improve the security of Android unlock patterns. Some of these suggestions include the use of strength meters during pattern selection [2, 25, 26], rearrangement of the grid points [28, 29], use of background images during pattern selection [31], modification of the pattern size to prevent various attacks [13, 18, 24, 30] , forcing users to choose certain points during pattern selection [12], or the use of Double Patterns [14]. However, all these suggestions have their drawbacks. For instance, increasing the grid size has been shown not to improve security [4] and strength meters are constrained by the inaccuracy of their underlying algorithms [15]. It is also unclear if methods that fundamentally change pattern entry, like Double Patterns [14] or Pass-O [28], will have widespread user support or adoption, despite security benefits.

Here, we propose using blocklists, which do not change the input interface, and have evidence of positive security effects, such as by Markert et al. [20] for PINs, Samuel et al. [22] for Knock Codes, and Forman and Aviv [14] for Double Patterns. Markert et al. [22] found that a small, enforcing blocklist would have large effects on PIN guessability, and that a blocklist of approximately 1000 PINs would properly balance usability and security. Forman and Aviv [14] found that small blocklists of first-pattern selection for Double Patterns had a similarly outsized effect on security, and Samuel et al. [22] found that blocklists significantly improve the security of Knock Codes. While our study similarly explores the security and usability of blocklists on smartphone authentication, it differs from the above studies by focusing on traditional, unmodified Android unlock patterns. In the end, we find that blocklists, even relatively small ones, can significantly improve the security of unlock patterns, inline with prior results [14, 20, 22].

3 Methodology

In the following, we describe our methodology. We start by outlining the design of the user study and giving a detailed description of the 6 treatments, and following, we discuss the recruitment process, limitations, and ethics.

3.1 Survey Structure

We conducted an online survey on Amazon Mechanical Turk (MTurk), and to ensure ecological validity of selecting and entering patterns on a mobile device, the survey was designed

to be taken on mobile browsers only, as checked via the user-agent. Our study was open to both pattern and non-pattern users, with pattern users free to select patterns they already use unless blocked as part of a blocklist treatment. Participants were assigned 1 of 6 treatments for selecting and recalling a pattern: 5 blocklist-enforcing treatments with blocklists of various sizes and 1 control treatment without any blocklist intervention. We will discuss those treatments in more detail in Section 3.2. The average time to complete the survey was 6 minutes and participants were compensated $1.00.

We will now outline the structure of the survey; for a detailed description, please refer to Section A in the Appendix.

1. *Informed Consent*: Participants were informed about the purpose, structure, and anticipated duration of the study as well as the compensation.
2. *Device Usage*: Participants were asked about the number of smartphones they use, the device brands, and their authentication methods. Details regarding device usage can be found in Table 7.
3. *Background Information*: Because we could not expect all participants to be familiar with Android patterns, we provided background information including how to create a valid pattern. We further showed them an image with the Android unlock pattern interface, but not an entered pattern to avoid priming.
4. *Practice*: Participants were asked to practice creating a pattern before proceeding, serving as a hands-on introduction to patterns. The patterns selected here are not used in our analysis.
5. *Instructions/Scenario-Priming*: After familiarizing with Android unlock patterns, participants were informed that they should now create a pattern that they would use to secure their primary smartphone. Participants were also informed that they would have to recall the pattern they selected and therefore, it would need to be both secure and memorable. Participants were additionally instructed not to write down or use any aids to help them remember their pattern. To proceed, participants had to confirm that they understood all of the mentioned instructions.
6. *Selection/Blocklist-Intervention*: Participants selected (and confirmed) a pattern as they would use to secure their primary smartphone. During selection, participants in the blocklist treatments saw the warning depicted in Figure 1 if they entered a disallowed pattern and were asked to select a different pattern.
7. *SUS*: After selecting a pattern, participants answered questions from the System Usability Scale (SUS) to determine their perceived usability of pattern selection.
8. *Post-Entry*: In addition to the SUS questions, participants were asked whether they felt they created a pattern that provides adequate security and whether it was difficult for them to select the pattern.
9. *Strategy*: To understand how users select and change their patterns, we asked participants that encountered a blocklist for their selection strategy prior to the warning and how their strategy changed after seeing it. Participants that did not encounter a blocklist were asked to imagine how their strategy would change if they encountered the blocklist warning.
10. *Recall*: Participants attempted to recall their pattern within 5 attempts.
11. *Security Comparison*: After recall, participants were asked about the security of patterns in general and in comparison to 3 other unlock methods: 4-digit, as well as 6-digit PINs, and alphanumeric passwords.
12. *Real World Usage*: To better understand whether the patterns created in the study would actually be used, we asked participants if they would select the same unlock pattern on their smartphones along with their reasons for that decision.
13. *Demographics*: Participants were asked to provide demographic information, such as age, identified gender, dominate hand, education, and technical background. We also included a second attention check question on this page. To ensure that the demographic backgrounds of the participants do not interfere with the rest of the study [21], we asked these questions at the very end of the study.
14. *Honesty*: Finally, we asked participants if they had honestly participated in the survey and followed instructions completely. We paid all participants who completed the study but discarded participants from the analysis if they indicated dishonesty at this point.

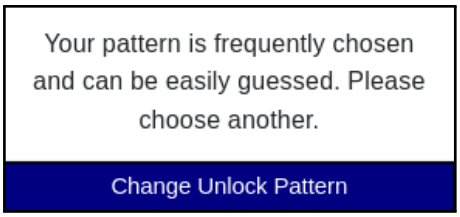


Figure 1: Blocklist warning used in the study.

3.2 Treatments

Participants were randomly assigned to either a control treatment or 1 of the 5 blocklist-enforcing treatments. To determine the common patterns to block, we combined data from von Zezschwitz et al. [31], Aviv et al. [4], Uellenbeck et al. [29], and Loge et al. [19], for a total of 4,637 patterns. Blocklists were generated by selecting patterns that appeared at least a certain number of times in the data set, e.g., at least 2 times for BL-2 (the largest blocklist with 581 patterns) or at least 32 times BL-32 (the smallest blocklist with 12 patterns).[1] The treatments are described below:

[1] To foster future research on this topic, we share the described blocklists. Please contact the authors for this purpose.

- **Control** ($n = 169$): Participants received no interventions when selecting a pattern.
- **BL-2** ($n = 166$)**:** The blocklist in this treatment comprised 581 patterns, with these patterns appearing at least twice in prior work.
- **BL-4** ($n = 172$)**:** The blocklist in this treatment comprised 239 patterns, with these patterns appearing at least 4 times in prior work.
- **BL-8** ($n = 161$)**:** The blocklist in this treatment comprised 105 patterns, with these patterns appearing at least 8 times in prior work.
- **BL-16** ($n = 165$)**:** The blocklist in this treatment comprised 54 patterns, with these patterns appearing at least 16 times in prior work.
- **BL-32** ($n = 173$)**:** The blocklist in this treatment comprised 12 patterns, with these patterns appearing at least 32 times in prior work.

Participants in the blocklist treatments received the warning message in Figure 1 when they selected a blocked pattern, which is based on the iOS blocklist warning [14, 20]. Blocklists were enforcing, i.e., could not be ignored, and participants were required to select a pattern that was not blocked.

3.3 Recruitment and Demographics

We recruited $n = 1006$ participants on Amazon Mechanical Turk (MTurk), after excluding 65 responses due to failed attention checks or dishonesty. As expected when recruiting from MTurk, our surveyed population was comprised primarily of younger (59 % between 18–34), male-identifying (62 % male, 36 % female, and 2 % other gender, or prefer not to say) participants with semi- or full college education (28 % some college or Associate's, 60 % Bachelor's or above). Table 1 depicts the full demographic information.

Table 1: Demographic information of participants.

	Male		Female		Other		Total	
	No.	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**
Age	624	62 %	367	36 %	15	1 %	1006	100 %
18–24	80	8 %	46	5 %	1	0 %	127	13 %
25–29	150	15 %	84	8 %	3	0 %	237	24 %
30–34	137	14 %	79	8 %	1	0 %	217	22 %
35–39	106	11 %	69	7 %	4	0 %	179	18 %
40–44	54	5 %	25	2 %	0	0 %	79	8 %
45–49	48	5 %	27	3 %	0	0 %	75	7 %
50–54	27	3 %	11	1 %	0	0 %	38	4 %
55–59	9	1 %	11	1 %	0	0 %	20	2 %
60–64	5	0 %	6	1 %	0	0 %	11	1 %
65+	8	1 %	9	1 %	0	0 %	17	2 %
Prefer not to say	0	0 %	0	0 %	6	1 %	6	1 %
Education	624	62 %	367	36 %	15	1 %	1006	100 %
Some High Sch.	0	0 %	2	0 %	0	0 %	2	0 %
High School	56	6 %	29	3 %	0	0 %	85	8 %
Some College	119	12 %	66	7 %	1	0 %	186	18 %
Trade	17	2 %	9	1 %	0	0 %	26	3 %
Associate's	51	5 %	44	4 %	1	0 %	96	10 %
Bachelor's	288	29 %	168	17 %	5	0 %	461	46 %
Master's	74	7 %	41	4 %	1	0 %	116	12 %
Professional	10	1 %	5	0 %	0	0 %	15	1 %
Doctorate	9	1 %	3	0 %	0	0 %	12	1 %
Prefer not to say	0	0 %	0	0 %	7	1 %	7	1 %
Background	624	62 %	367	36 %	15	1 %	1006	100 %
Technical	266	26 %	87	9 %	1	0 %	354	35 %
Non-Technical	335	33 %	265	26 %	4	0 %	604	60 %
Prefer not to say	23	2 %	15	1 %	10	1 %	48	5 %

3.4 Limitations

Our study has a number of limitations. Due to the nature of online surveys, it is not possible to tell whether participants fully and completely followed the instructions provided in the survey. We tried to mitigate this by including 2 attention check questions in the survey and asking participants whether or not they answered honestly, highlighting that they would be paid irrespective of their answer. Additionally, we reviewed all participant responses and removed participants from our analysis whose responses were inconsistent. As with other studies, participants on MTurk tended to be younger and more educated. We do not make any claims about our results being representative of the general population. As our study was relatively short, the recall rates reflect short-term memorability of unlock patterns; future work is needed to explore long-term memorability of these patterns. However, this approach has been used with a lot of success by many other researchers in the community to study mobile authentication [3, 4, 14, 19, 20, 29].

This survey may have been participants' first exposure to unlock patterns (27 % of participants were pattern users), and as a result, the non-pattern users' selection may vary in a real-world setting. To test for this, we asked non-pattern users if they would use the pattern they created to secure their primary smartphone. The results show that 40 % would use the pattern they created, 26 % were unsure and 34 % would not. Most participants who indicated that they were unsure or would not use their pattern argued that they would use it, had it not been recorded in the survey. This suggests that the patterns of the participants who have not previously used this unlock method closely match up to pattern selection in the real world.

3.5 Ethical Considerations

The study was approved by our Institution's Review Board (IRB), and participants were fully informed about the purpose and structure of the study. All participants were paid regard-

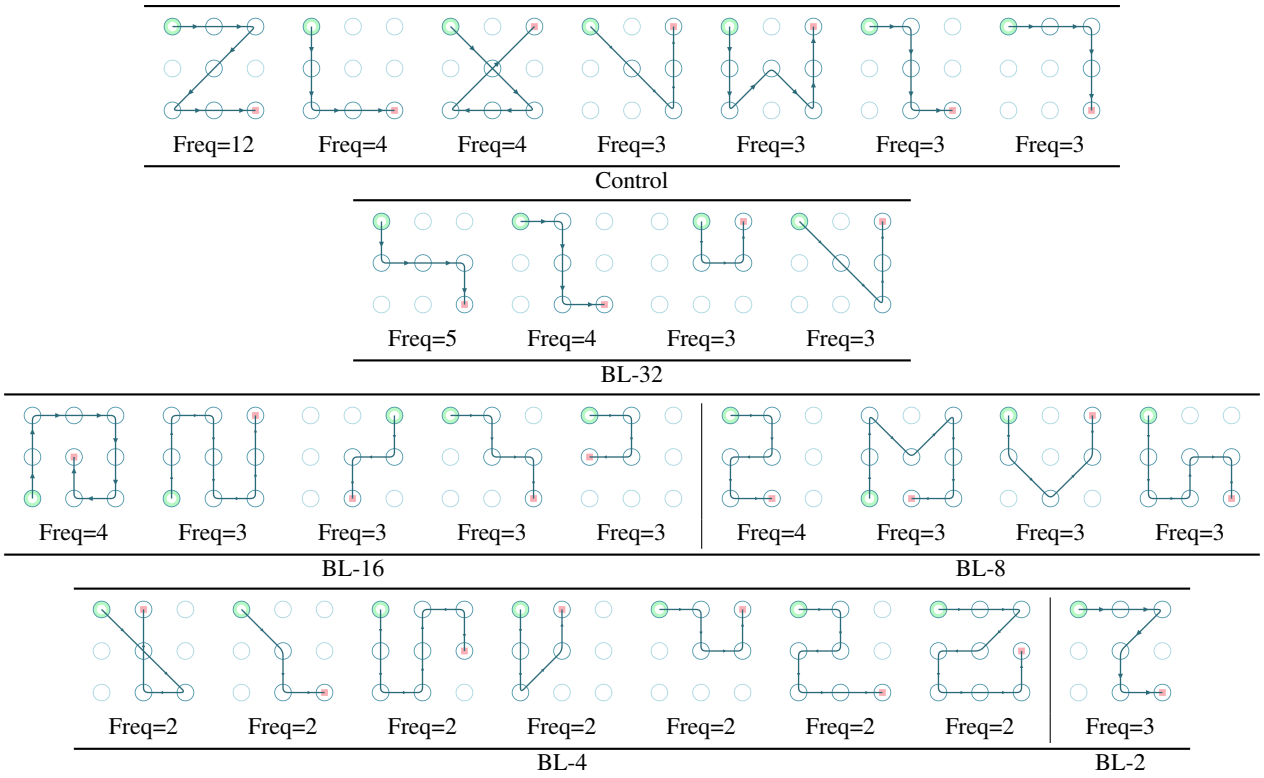


Figure 2: Most frequent patterns observed in treatments. Green circles depict start of a pattern, red squares indicate the end.

less of the quality of their submitted data. This includes cases where we removed submissions from our analysis for failing attention checks as well as situations where participants indicated dishonesty in answering the survey questions.

Another aspect to consider is the risk associated with the login information participants share with us through the study. As described earlier, some participants said they would use the unlock patterns they created in the study and others even confirmed that they do use the very same pattern to secure their smartphone. While a targeted attacker could potentially use this information to harm users, the implied risk is minimal. There is no identifiable connection from the selected unlock patterns to individual participants. On the other hand, this research offers much benefit as the outcomes of improved blocklists assist future selection of Android unlock patterns.

4 Features of Collected Patterns

In this section, we discuss pattern properties including the most frequent patterns, lengths of the patterns as well as common start and end points of patterns selected by participants across treatments. Figure 2 shows the most frequent patterns selected across the different treatments. The start point of each pattern is a green circle while the end point is a red square. An arrow indicates the direction of the pattern from the start point to the end point.

The most common patterns in the control treatment directly depict letters such as Z ($n = 12$), L ($n = 4$), and W ($n = 3$), along with patterns that resemble letters such as X ($n = 4$) or V ($n = 3$). The latter is also popular in the BL-32 treatment along with a small U ($n = 3$), but the most popular are 2 patterns which start in the upper left, move through the central point, and end in the lower right. In BL-16, flipped letters including G ($n = 4$), U ($n = 3$) and the number 2 ($n = 3$) are more common. The number 2 ($n = 5$) is the most frequent pattern in BL-8, followed by the letter V ($n = 3$). We also observe modifications to letters in this treatment, including addition of lines to the letter M ($n = 3$). Patterns in the BL-4 treatment are notably more diverse, with the common patterns only appearing twice at most, including more advanced modification of letters such as U ($n = 2$), Z ($n = 2$), and number 2 ($n = 2$). Similarly, patterns in BL-2 are more diverse, with a more advanced modification to letter Z ($n = 2$) being the only pattern appearing at least twice.

Many patterns in the control treatment appear to use shapes including numbers or letters in their exact form factor, while shapes are altered by flipping, mirroring or adding extra lines in the blocklist treatments. This is further confirmed through our qualitative analysis of users' pattern selection strategies: most participants initially select their patterns based on shapes such an initial of their name for memorability, but add complexity, such as extra lines, when they encounter a blocklist.

Table 2: Properties of selected patterns.

Treatment	Patterns No.	Unique Patterns No.	Unique Patterns %	Blocklist Hits No.	Blocklist Hits Average	Participants with Hits No.	Participants with Hits %	Length Average	Length Std. Dev.	Stroke Length Average	Stroke Length Std. Dev.
Control	169	130	77 %	0	0.00	0	0.0 %	6.1	1.5	5.4	1.7
BL-32	166	142	86 %	41	0.25	32	19.3 %	6.1	1.5	5.4	1.7
BL-16	172	151	88 %	98	0.57	61	35.5 %	6.1	1.7	5.4	1.8
BL-8	161	141	88 %	129	0.80	66	41.0 %	6.0	1.6	5.4	1.7
BL-4	165	158	96 %	202	1.22	86	52.1 %	6.3	1.7	5.7	1.9
BL-2	173	155	90 %	368	2.13	127	73.4 %	6.2	1.5	5.4	1.6
Total	1006	724	72 %	838	0.83	372	37.0 %	6.1	1.6	5.5	1.7

Figure 3: Frequency of start and end points. The top row shows the start points while the bottom shows the end points.

Figure 3 shows the most common start and end points for patterns, with the top row depicting start points and the bottom row depicting end points. As reported in prior work [3, 4, 29], most patterns start in the top left corner of the grid and end in the bottom right. However, this becomes less prevalent for patterns selected in blocklist treatments, with 36.0 % to 44.6 % of these patterns starting in the top left corner, instead of 49.1 % in the control treatment. For the end points, 15.7 % to 22.3 % of patterns in the blocklist treatments end in the bottom right corner compared to 32.5 % in the control group. While there was no significant difference in starting at the top left corner, a *chi*-square test showed significant difference in ending at the bottom right corner ($\chi = 17.65$, $p < 0.01$) across treatments. This suggests that blocklists likely pushed participants to change their end points more as compared to their start points when encountering a blocklist.

We also considered the lengths of patterns, both in terms of number of contact points used (i.e., *length*) and the length of the strokes within the pattern (i.e., *stroke length*) (see Table 2). The stroke length is calculated by taking the Cartesian difference with the origin mapped to the center point and unit distances between points. We find no significant differences for length ($f = 0.639, p = 0.66$) or stroke-length ($f = 0.937, p = 0.45$) between treatments. Participants select patterns of similar lengths, but varied other properties after encountering blocklists.

Finally, we compared the number of unique patterns across different treatments. As shown in Table 2, the percentage of unique patterns selected by participants increases with the blocklist size. While only 77 % of the patterns were unique in the control treatment, 90 % were in BL-2, the largest blocklist size, and even 96 % in the second largest blocklist BL-4. We performed a χ^2 test on the prevalence of unique patterns, finding there is a significant difference ($\chi = 11.04$, $p = 0.05$). Post-hoc analysis (Bonferoni-corrected) revealed that only BL-2 and Control were significantly different, where BL-2 had the highest rate of unique patterns.

5 Security Analysis

In this section, we describe the security analysis of patterns selected with and without a blocklist. First, we introduce the the attacker model for a perfect knowledge and simulated attacker, and then we discuss the success rates of the 2 guessing attacks. Lastly, we discuss an analysis of selecting a blocklist size that balances the security and usability of patterns.

Attacker Model. We make a number of assumptions for our attacker model. Foremost, the attacker is generic and does not have additional information about individual users to perform a targeted attack. Such an attacker could use tailored techniques, for example, shoulder surfing [5, 8, 23] or smudge attacks [6], which may increase the success rate for a given victim, but be less successful in general.

We also consider 2 variations of the generic attacker, a perfect knowledge and a simulated attacker. The perfect knowledge attacker provides an upper bound performance of the generic attacker since it assumes the attacker knows the exact distribution of frequencies of patterns, and always guesses the next most frequent pattern. On the other hand, a simulated attacker utilizes a set of training data to guess an unknown set of the authentication. For the simulated attacker, we also assume that the attacker has knowledge of the blocklist and optimizes the guessing order by skipping patterns which could not be selected. This is because an attacker would have access to the best training material, including the blocked patterns. For the perfect knowledge attacker, this assumption is always implied as the attacker is aware of the distribution.

5.1 Perfect Knowledge Attacker

The perfect knowledge attacker results are presented in Table 3. To control the different sizes of our treatment groups and allow for a fair comparison, we randomly down-sampled all larger data sets to 161, i.e., the size of the BL-8 treatment. For the strength estimations of the perfect knowledge attacker, we use two metrics, β-*success-rate* and α-*guesswork*, as defined by Bonneau [9].

First, for an attacker that is limited in the number of guesses as is the case with unlock patterns, β-success-rate describes the percentage of the dataset guessed after β guesses. Reported as λ_β in Table 3, it is evident that blocklists greatly reduce the success rate of such a throttled attacker. BL-4, the second largest blocklist size, appears to reduce the attacker performance the most across the scenarios we investigated. After 3 guesses, BL-32 (the smallest blocklist) reduces the attacker performance from 13.1 % down to 9.0 % of patterns successfully guessed, as compared to the control treatment. BL-4 reduces the attacker performance even further, with only 4.6 % of patterns guessed after 3 attempts. After 10 guesses, BL-32 reduces the attacker performance from 22.9 % to 18.0 % compared to the control group. BL-4 further reduces the attacker success rate to only 10.9 % of patterns guessed after 10 attempts. After 30 guesses, the attacker can guess 41.1 % of patterns in the control group, but only 33.1 % in BL-32 and 22.3 % in BL-4. This suggests that even small blocklists can improve the security of user-selected patterns.

The second metric, α-guesswork, measures an attacker who is not constrained by the number of attempts to guess an authentication, otherwise known as an *unthrottled attacker*. Using bits of entropy, it measures how much "work" is required to guess an α fraction of the data set. A higher entropy implies more work for the attacker and ultimately shows the authentication is stronger. These results are indicated by $\widetilde{G}_\alpha$ in Table 3. Across all cases, the attacker is less successful when guessing patterns in the blocklist treatments compared to the control group. Just like in the throttled setting, patterns selected in the BL-4 treatment are stronger, with the α-guesswork being higher compared to the other groups for all guessing scenarios evaluated. When guessing 50 % of the data, BL-32, the smallest blocklist increases the guessing entropy by 0.33 compared to the control treatment. BL-4 further increases the entropy by 0.68 as compared to control. This again advocates that blocklists increase security of unlock patterns.

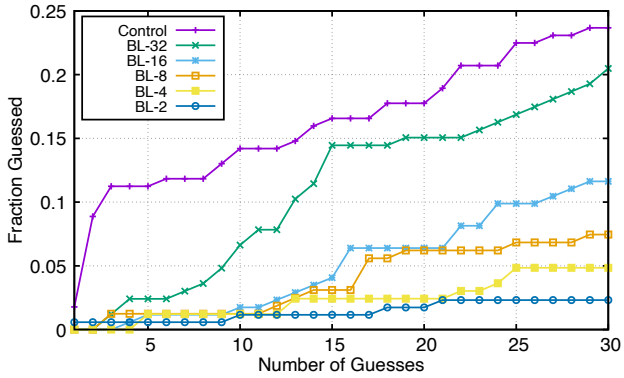
Table 3: Guessing metrics for a perfect knowledge attacker.

	Throttled Attack (%)			Unthrottled Attack (Bits)			
Treatment	λ_3	λ_{10}	λ_{30}	H_∞	$\widetilde{G}_{0.1}$	$\widetilde{G}_{0.3}$	$\widetilde{G}_{0.5}$
Control	13.1 %	22.9 %	41.1 %	3.75	4.66	6.00	6.93
BL-32	9.0 %	18.0 %	33.1 %	5.01	5.82	6.65	7.26
BL-16	7.3 %	15.6 %	29.9 %	5.33	6.04	7.00	7.45
BL-8	8.0 %	17.1 %	31.9 %	5.33	5.89	6.81	7.33
BL-4	4.6 %	10.9 %	22.3 %	6.33	6.64	7.34	7.61
BL-2	5.1 %	13.1 %	27.9 %	5.75	6.31	7.00	7.43

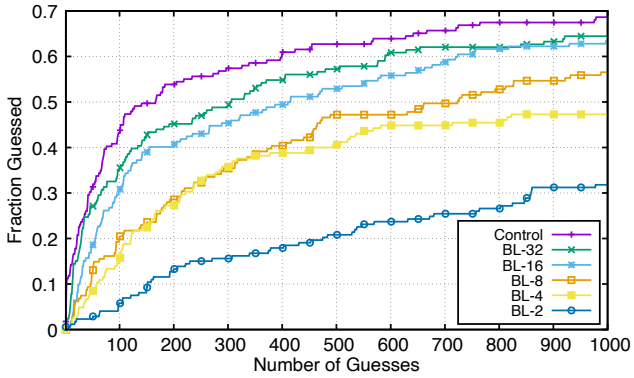
5.2 Simulated Attacker

A simulated attacker guesses a set of unknown authentications based on a set of training data. Using published data of Android patterns from von Zezschwitz et al. [31], Aviv et al. [4], Uellenbeck et al. [29], and Loge et al. [19], we first created a training data set where we ordered the patterns by their frequency of occurrence, starting from the most common patterns. Our training set consisted of a total of 4,637 patterns of which 581 are unique. In cases where multiple patterns had a similar number of occurrences, we used a Markov Model to order them based on their probability of occurrence. Using data of all possible unlock patterns from Aviv et al. [4], we trained our model to compute the transition probabilities, using Laplace smoothing to ensure no zero probability transitions existed for valid transitions not appearing in the training data. Using the Markov Model once more, we extended our initial training set by adding all other possible Android unlock patterns based on their likelihood of occurrence. Our final training data set was comprised of 389,112 patterns, the total possible number of Android unlock patterns.

Figure 4a and Figure 4b show the results of a simulated guess for up to 30 and 1000 patterns respectively. As can be seen from both graphs, blocklists reduce the fraction of patterns guessed. After 30 guesses, the simulated attacker can guess 23.7 % of patterns in the control treatment, 20.5 % in BL-32, 11.6 % in BL-16, 7.5 % in BL-8, 4.8 % in BL-4, and 2.3 % in BL-2. After 1000 guesses, the attacker can guess 68.7 % of patterns in the control group, 64.5 % in BL-32,



(a) Success rate for up to 30 guesses.



(b) Success rate for up to 1000 guesses.

Figure 4: Success rates of a simulated attacker.

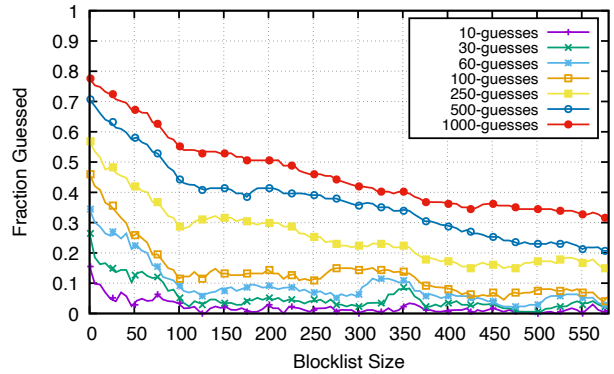


Figure 5: Effect of blocklist size on simulated guessing.

63.4 % in BL-16, 56.5 % in BL-8, 47.3 % in BL-4, and 31.8 % in BL-2. Hence, we can conclude that similar to perfect knowledge guessing, even the smallest blocklists reduce the simulated attacker's success rate.

5.3 Appropriate Blocklist Size

While a large blocklist is beneficial for security, it is important to have an appropriate blocklist size to limit negative effects on the user experience. As shown in Table 2, chances of users encountering a blocklist are higher as the blocklist size increases. Hence, we now discuss an appropriate blocklist size to balance the security and usability of unlock patterns.

Our experiment allowed us to collect not just the final patterns but also all other patterns selected by participants that were rejected due to a blocklist. With knowledge of each participant's pattern selection attempts, we simulated different blocklist sizes to determine the pattern they would have selected given a certain blocklist size. Finally, we performed a simulated guessing attack to determine the fraction of patterns that would be guessed for the different simulated blocklist sizes after a varied number of guessing attempts.

Figure 5 shows our simulation results. Initially, a lot of patterns can be guessed when there is no blocklist in place, i.e., when the blocklist size is 0. As the blocklist size increases, the fraction of patterns guessed also decreases through a series of dips and peaks, caused by participants settling again on popular patterns after encountering a blocklist warning on their previous choice. By entering the first dip for instance, the attacker is most disadvantaged as it is no longer possible to solely rely on guessing first choice patterns; but more and more second choice patterns need to be considered as well. Ultimately, the blocklist restricts all first choices and therefore, the attacker can now guess popular second choices which results in a peak.

These series of dips and peaks suggest that a properly sized blocklist should be based on one of the dips as this is where the attacker is most disadvantaged. To achieve this while having minimal effect on usability, the first dip for 30 to 60 guesses that translates to a blocklist size of about 100 patterns appears to be the most ideal. This is most similar to the BL-8 treatment which blocked 105 patterns.

6 Pattern Selection Strategies

In this section, we discuss the strategies that participants used to select patterns when encountering a blocklist. Participants were asked about both their initial strategy for selecting a pattern and how that strategy changed when they encountered a blocklist. Those who did not encounter a blocklist, were asked to imagine how they would change strategies. We qualitatively coded a random sub-sample of 309 responses (about a third of the sample space), split comparably across treatments. Two coders independently coded the responses and met to collaboratively review discrepancies until agreement was met. We settled on 11 primary codes that describe participants' selection strategies. A full description of the codes can be found in Table 5 in Appendix B.

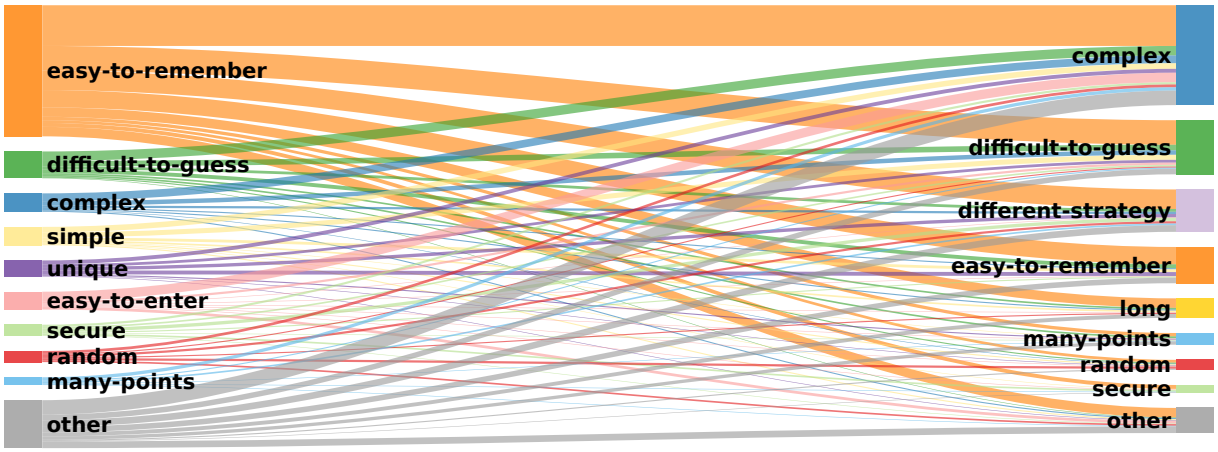


Figure 6: Changes to pattern selection strategies upon encountering a blocklist.

Initial Strategies. To understand the initial selection strategies, participants were asked about how they selected their unlock pattern, with participants that encountered a blocklist specifically asked about their strategy prior to encountering the blocklist warning. The vast majority of participants indicated selecting a pattern that would be easy to remember (70.6 %). This matches inquiries from prior work [14].

Other common strategies mentioned by participants included patterns that would be "difficult to guess" (15.1 %) and "complex" (10.0 %). There was a roughly equal split between participants who valued complexity and simplicity in their patterns, with another 10.0% of responses being tagged as "simple." 9.1 % of participants indicated that they chose their pattern to be "unique" or "uncommon" while 8.7 % mentioned selecting a pattern that would be "easy to enter".

These results indicate that prior to encountering a blocklist, most users are more concerned about selecting a pattern that would be easy to remember rather than secure. This has also been demonstrated in other studies whereby most users choose convenience over security or privacy [17].

Post-Blocklist Strategies. Participants that encountered a blocklist warning were asked how their strategies changed upon encountering the blocklist while those that did not were asked to describe how they imagine their strategies would change upon encountering such a warning. The greatest percentage of participants indicated choosing "complex" patterns (49.5 %) after encountering a blocklist, while 28.2 % changed their patterns to be "difficult to guess". A further 20.7 % of participants indicated that they would change their strategy in some way but did not specify exactly how they would do so.

Only 17.8 % of participants chose "easy-to-remember" patterns following a blocklist warning, a significant reduction compared to participants that used this strategy prior to encountering a blocklist. About 13.9 % of participants' responses were tagged as "long" or "many-points", meaning that they wanted their patterns to cover many contact points for a longer pattern, indicating a desire for complexity or a pattern that is harder to guess. A small percentage (5.5 %) of participants stated that they chose their pattern at random after the blocklist encounter.

These results show the positive security effects of blocklists, with a majority of users indicating using strategies that are more security minded, either making their patterns more complex or harder to guess.

Changes of Strategy. Figure 6 shows how participants' pattern selection strategies changed after they encountered a blocklist, with their strategies prior to a blocklist on the left and their strategies after encountering a blocklist on the right.

The most significantly changed strategies were "easy-to-remember" and "complex", with the number of participants who selected their pattern to be easy to remember decreasing by about 74.8 %. The participants who selected complex patterns increase by 393.5 %, after encountering a blocklist. The number of participants whose responses indicated security in general (complex, difficult-to-guess, long, secure) increased by 190 %. Additionally, before encountering a blocklist, only 4.2 % of participants indicated that they wanted their patterns to be long. In contrast, 13.9 % of participants increased their pattern length after encountering a blocklist warning.

Our results suggest that when users are not primed to think about the security of their patterns, they tend to prefer memorability and convenience. However, after they encountered a blocklist, the most common strategies were to make their patterns complex (49.5 %) and difficult to guess (28.2 %). This shows that blocklists can meaningfully encourage users to consider security just as much as they consider convenience when selecting patterns to secure their smartphones.

Table 4: Usage statistics for the control and the 5 blocklist treatments.

	Control	BL_32	BL_16	BL_8	BL_4	BL_2	Hit BL	No BL	Total
Mean Selection Time	13.64s	13.41s	16.67s	19.27s	25.52s	34.24s	34.50s	12.31s	20.52s
Median	7.38s	9.12s	12.34s	13.88s	17.48s	26.70s	27.99s	7.74s	13.38s
Standard Deviation	26.91s	12.17s	15.98s	17.04s	25.25s	29.23s	24.14s	18.34s	23.27s
Mean Entry Time	1.53s	1.46s	1.53s	1.73s	1.87s	1.79s	1.75s	1.59s	1.65s
Median	1.27s	1.19s	1.33s	1.46s	1.53s	1.62s	1.52s	1.32s	1.40s
Standard Deviation	1.10s	0.94s	0.83s	1.00s	1.35s	0.91s	1.04s	1.04s	1.04s
Mean Recall Attempts	1.33	1.35	1.27	1.35	1.52	1.52	1.53	1.31	1.39
Median	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Standard Deviation	0.87	0.82	0.64	0.78	1.03	1.03	1.00	0.79	0.88
Recall Success Rate	100.00%	99.55%	100.00%	99.54%	100.00%	99.62%	99.82%	99.76%	99.78%
Mean SUS Score	78.64	78.77	78.01	76.96	76.47	71.62	71.40	79.84	76.72
Median	82.5	80.0	80.0	80.0	77.5	75.0	72.5	82.5	77.5
Standard Deviation	17.37	16.51	16.47	16.84	16.80	17.82	17.74	15.95	17.12

7 Usability

In this section, we discuss the usability of patterns selected across treatments. We begin by discussing the amount of time participants took to select and enter their patterns followed by short-term recall rates across treatments. Afterwards, we discuss System Usability Score (SUS) before reporting on a series of Likert-based responses regarding usability.

Selection Time. Our study recorded the amount of time participants took to select and enter their patterns. As can be seen in Table 2, participants in the control group took on average 13.64 seconds to select a pattern, compared to 34.24 seconds on average in the largest BL-2 treatment. The 151 % increase in selection time is likely due to users encountering the blocklist multiple times as well as the extra time needed to develop more complex patterns. Among the smaller blocklist treatments, the average selection time varied by a few seconds, with participants requiring on average 5.63 more seconds to select a pattern in BL-8 compared to the control group. Using a one-way analysis of variance (ANOVA), we find significant difference ($f = 22.06$, $p < 0.05$) for selection time across treatments. By performing a post-hoc pairwise analysis (with Holm-Sidak correction), we find that the control treatment is not significantly different from the small blocklist treatments i.e. BL-32 ($p = 0.99$), BL-16 ($p = 0.99$) and BL-8 ($p = 0.76$) but significantly differs from the large blocklist groups ($p < 0.05$) i.e. BL-4 and BL-2. Further, Cohen's effect size values suggest a medium effect size between the control group and BL-4 ($d = 0.46$), but a fairly large effect size between the control group and BL-2 ($d = 0.73$). These results indicate that larger blocklists can significantly increase the time used to select a pattern, showing the need to appropriately size blocklists to preserve the usability of unlock patterns.

Entry Time. The average entry times remained mostly unaffected by the blocklists. In the control, participants took on average 1.53 seconds. The only notable changes can be seen for the large blocklist treatments where entry times rose marginally to 1.87 seconds for participants in the BL-4 treatment and 1.79 seconds for BL-2. A one-way ANOVA found significant difference ($f = 4.10$, $p < 0.05$) for entry time across treatments. However, after performing a post-hoc pairwise analysis (with Holm-Sidak correction) we do not find significant difference between any of the treatments, suggesting that blocklists have limited impact on entry time of patterns.

Recall. The vast majority of participants were able to recall their patterns later in the survey as shown in Table 4 regardless of their treatment. Recall rates, albeit short-term, were not significantly different across treatments nor between those that hit and those that did not hit a blocklist. However, the average number of attempts needed to recall their patterns did vary across treatments. In the control group, users needed 1.33 attempts while BL-2 treatment participants required 1.52 attempts on average. The users who did not hit a blocklist within any treatment took 1.31 attempts and those who did took 1.53 attempts on average. While these results suggest that patterns selected after encountering a blocklist tend to be slightly less memorable compared to those selected without a blocklist, we do not find any significant difference ($H = 9.40$, $p = 0.09$) across the treatments using the Kruskal-Wallis test. Note the recall rates captured in our experiment were short-term due to our focus on the immediate impact of blocklists; exploring long-term recall is a promising area of future work.

Usability Perceptions. We used the System Usability Scale (SUS) to measure the perceived usability of the pattern scheme in presence of different blocklists. As shown in the last row of Table 4, the SUS scores are acceptable across all

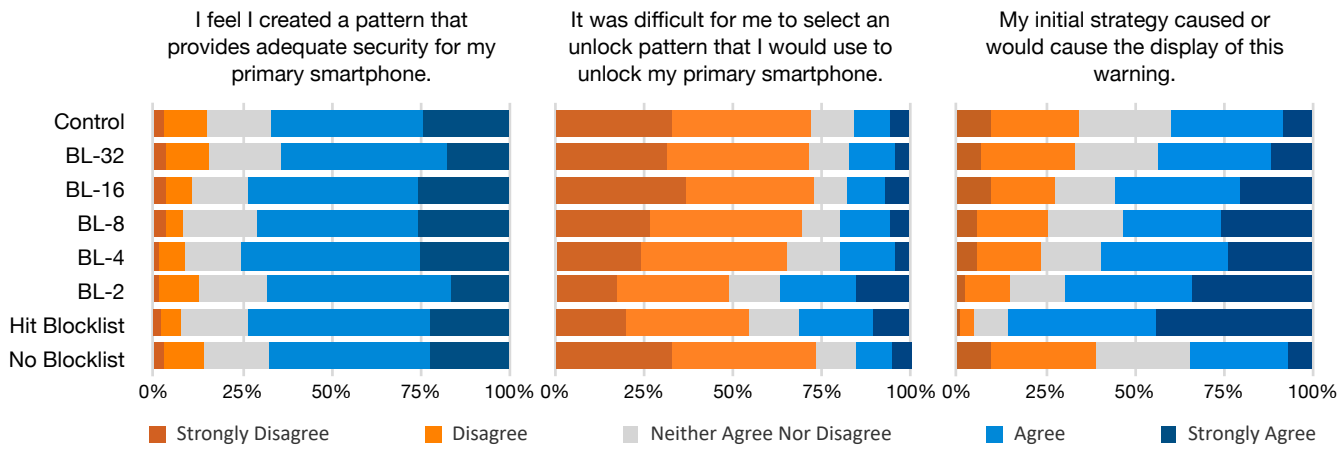


Figure 7: Agreement with Likert-scale questions relating to security and usability perceptions of unlock patterns.

the treatments. Apart from BL-2 where users recorded an SUS score of 71.6, all other blocklist treatments had SUS scores that were comparable to the control treatment, ranging from 76.5 to 78.6. We anticipate the lower but acceptable SUS score in BL-2 was due to users getting frustrated due to the large blocklist size. Therefore, it is important to use an appropriate blocklist size in order to have minimal impact on the usability of unlock patterns.

Security vs. Usability Tradeoffs. Participants were asked a series of Likert-scale questions on the perceived security and usability of their patterns (cf. Figure 7). First, participants were asked if they felt that they "created a pattern that provides adequate security" for their primary smartphone. Across all treatments, ~70 % of participants agreed that they chose a secure pattern regardless of blocklist encounters. This could be due to social desirability bias, where participants over-report the security of their patterns to seem more favorable in a security-focused study. However, a Mann-Whitney U test showed significant difference ($U = 110674.5$, $p < 0.05$) in perceived security for participants that did and did not encounter a blocklist, suggesting that encountering a blocklist marginally increases ($\eta^2 = 0.003$) users' perception of the security of their patterns.

Participants were also asked if it was difficult to select an unlock pattern that they would use to unlock their primary smartphone. The percentage of participants who agreed with the statement increased as the blocklist size grew, meaning that strict blocklists made it harder for people to select usable patterns. BL-2 (20.0 %) and BL-4 (37.0 %) treatments had the largest percentage of participants agreeing with this statement. Using a Mann-Whitney U test, we find that participants that encounter a blocklist think it is more difficult ($U = 90388.5$, $p < 0.05$) to select a pattern compared to participants that do not. We anticipate that this small increase ($\eta^2 = 0.038$) is likely caused by user frustration with large blocklists, further reinforcing the need to appropriately size blocklists.

When prompted about their agreement with the statement "my initial strategy caused the display of this warning" over 80 % of participants that encountered a blocklist agreed. In contrast, less than 35 % of participants that did not hit a blocklist agreed that their initial strategy would cause the display of the warning. The control group participants were split roughly evenly with slightly more people agreeing with the statement. As the blocklist size increased, more participants agreed that their strategy caused the warning, with 70 % of BL-2 participants agreeing. A Mann-Whitney U test showed significant difference ($U = 43562.0$, $p < 0.05$) in agreement with the statement for those who encountered the blocklist versus those who did not. This suggests that after encountering a blocklist, users are 27.8 % more likely ($\eta^2 = 0.278$) to think critically about the security of their patterns

Our findings on the usability of Android patterns with blocklists seem to support our hypothesis. While large blocklists do increase users' perception of the security of their patterns, they also make patterns less memorable and less usable. Moderate blocklists such as BL-8 and BL-16 seem to improve security without the huge usability trade-off incurred by BL-2. This further shows the need to select an appropriate blocklist size to avoid user frustration during pattern selection.

8 Discussion

Android unlock patterns continue to be a popular mobile authentication mechanism, 27 % of respondents in our study use them, matching similar reported usage in prior studies [17, 20]. This makes Android patterns the second most commonly used authentication on mobile devices after PINs. However, despite their popularity, patterns are comparatively less secure than both PINs and passwords [4, 20, 29], and have remained largely unchanged since they were first launched in 2008, with no significant security updates.

Our work suggests that the usage of blocklists, even quite small in size, can have dramatic improvements on the security of user-chosen unlock patterns. The blocklist warnings also primed participants to be more security conscious of their pattern choices, and had limited impact on short-term recall and entry times. Our results indicate that a blocklist with around 100 patterns would balance the security and usability needs sufficiently and could be deployed quickly and efficiently with minimal changes to Android's existing pattern interface.

Compared to most suggestions proposed to improve the security of unlock patterns such as rearrangement of points on the grid [28], use of strength meters [2, 25, 26] or providing guidance during selection [12], blocklists require the least updates to the simple interface that makes patterns so popular. In fact, existing warnings such as the one in use on Apple iOS can easily be adapted. Further, while blocklists have already been shown to improve security on other mobile authentication schemes such as PINs [20] and Knock Codes [22], suggestions such as increasing the grid size have proven not to have meaningful security benefits [4]. While proposals such as Double Patterns [14] improve security, it remains unclear if they will be widely adopted because unlike blocklists, they alter both the selection and entry procedure of unlock patterns.

Our qualitative results demonstrated how users do not have a good sense of the security of their pattern choices, with most users (even those that selected easily guessable patterns) indicating their patterns to be secure. While this may be due to social desirability bias, we do observe that encountering a blocklist forces users to think about security of their patterns, with users resorting to patterns that are either complex or difficult to guess. In contrast, most users are primarily concerned about memorability of their patterns prior to encountering a blocklist. This suggests that the usage of blocklists can force users to consider security when selecting unlock patterns.

The biggest challenge with deployment of blocklists is asking participants to update their pattern if the one they currently use is on the blocklist. Research on password reuse notifications may be of benefit in solving this problem. For example, Golla et al. [16] investigate different notifications for password reuse which could be adapted to encourage participants to update their pattern to one not on the blocklist, including forcing a password reset. Since non-enforcing blocklists have been shown to have limited security benefits on user-chosen PINs [20], we recommend using enforcing blocklists that would force users to select patterns more diversely.

Long term memorability of patterns selected in the presence of blocklists could pose another challenge to the adoption of blocklists on Android patterns. While short-term recall times and attempts only varied marginally for participants in the blocklist treatments compared to the control group, our study design did not allow us to measure recall over an extended duration of time, which can be explored further in future work. However, similar approaches have been successfully used in prior work [4, 20, 22] in the security community. Additionally, blocklists have successfully been used on other mobile authentication schemes such as PINs [20].

While we primarily focus on a simulated and perfect knowledge attacker for our analysis, further work is needed to determine how other attackers including a targeted attacker would perform in guessing patterns, particularly if they are aware of the blocklist used. While we observe positive change of strategies from simple to complex after encountering a blocklist, this very information could further improve the guessing performance of an informed attacker. On the other hand, this change of strategy is likely to make it harder for attacks such as shoulder surfing [5].

Future work may also analyze pattern strategies used by participants in real time in order to provide more tailored blocklist warnings. For instance, if a user selects a shape such as a letter, the blocklist warning could inform the user that their pattern is a letter and can therefore be easily guessed. Other areas for more work include investigating whether different blocklists are needed for different communities. Our blocklists were constructed using common patterns observed in prior work, with these patterns primarily collected from users in Western countries. This might explain the reason for most users starting their patterns from the upper left corner as Western writing begins from the top left. Other work could also explore whether the blocklists would need to be updated over time as this was outside the scope of our study.

9 Conclusion

In this paper, we studied the security and usability of blocklists on user-selected unlock patterns, a feature currently unavailable on Android but used by Apple's iOS to improve the security of user-selected PINs. We conducted an online survey where $n = 1006$ participants selected patterns across 6 treatments: a control treatment and 5 blocklist-enforcing treatments. We find that even small blocklists improve the security of unlock patterns. For a simulated attacker that must guess patterns based on some training data, the attacker's performance is reduced from 24 % to 20 % of patterns successfully guessed after 30 guesses; the largest blocklist reduces the attacker's performance further down to only 2 % after a similar number of guessing attempts. For usability, blocklists had minimal impact on short-term recall rates and entry times, with SUS scores indicating good usability when selecting patterns even in the presence of a blocklist. From our results, we recommend a blocklist size of about 100 patterns to balance the security and usability of patterns. Adding this feature to the existing implementation of Android patterns can be done easily and does not require changes to the original interface.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1845300. Further support was received through the research training group "Human Centered Systems Security" sponsored by the state of North Rhine-Westphalia, Germany, and the German Research Foundation (DFG) within the framework of the Excellence Strategy of the Federal Government and the States – EXC 2092 CASA – 390781972.

References

[1] Daniel Amitay. Most Common iPhone Passcodes, June 2011. http://danielamitay.com/blog/2011/6/13/most-common-iphone-passcodes, as of June 10, 2021.

[2] Panagiotis Andriotis, Theo Tryfonas, and George Oikonomou. Complexity Metrics and User Strength Perceptions of the Pattern-Lock Graphical Authentication Method. In *Conference on Human Aspects of Information Security, Privacy and Trust*, HAS '14, pages 115–126, Heraklion, Crete, Greece, June 2014. Springer.

[3] Panagiotis Andriotis, Theo Tryfonas, George Oikonomou, and Can Yildiz. A Pilot Study on the Security of Pattern Screen-Lock Methods and Soft Side Channel Attacks. In *ACM Conference on Security and Privacy in Wireless and Mobile Networks*, WiSec '13, pages 1–6, Budapest, Hungary, April 2013. ACM.

[4] Adam J. Aviv, Devon Budzitowski, and Ravi Kuber. Is Bigger Better? Comparing User-Generated Passwords on 3x3 vs. 4x4 Grid Sizes for Android's Pattern Unlock. In *Annual Computer Security Applications Conference*, ACSAC '15, pages 301–310, Los Angeles, California, USA, December 2015. ACM.

[5] Adam J. Aviv, John T. Davin, Flynn Wolf, and Ravi Kuber. Towards Baselines for Shoulder Surfing on Mobile Authentication. In *Annual Conference on Computer Security Applications*, ACSAC '17, pages 486–498, Orlando, Florida, USA, December 2017. ACM.

[6] Adam J. Aviv, Katherine Gibson, Evan Mossop, Matt Blaze, and Jonathan M. Smith. Smudge Attacks on Smartphone Touch Screens. In *USENIX Workshop on Offensive Technologies*, WOOT '10, pages 1–7, Washington, District of Columbia, USA, August 2010. USENIX.

[7] Adam J. Aviv, Benjamin Sapp, Matt Blaze, and Jonathan M. Smith. Practicality of Accelerometer Side Channels on Smartphones. In *Annual Computer Security Applications Conference*, ACSAC '12, pages 41–50, Orlando, Florida, USA, December 2012. ACM.

[8] Adam J. Aviv, Flynn Wolf, and Ravi Kuber. Comparing Video Based Shoulder Surfing with Live Simulation and Towards Baselines for Shoulder Surfing on Mobile Authentication. In *Annual Conference on Computer Security Applications*, ACSAC '18, pages 453–466, San Juan, Puerto Rico, USA, December 2018. ACM.

[9] Joseph Bonneau. The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords. In *IEEE Symposium on Security and Privacy*, SP '12, pages 538–552, San Jose, California, USA, May 2012. IEEE.

[10] Joseph Bonneau, Sören Preibusch, and Ross Anderson. A Birthday Present Every Eleven Wallets? The Security of Customer-Chosen Banking PINs. In *Financial Cryptography and Data Security*, FC '12, pages 25–40, Kralendijk, Bonaire, February 2012. Springer.

[11] Seunghun Cha, Sungsu Kwag, Hyoungshick Kim, and Jun Ho Huh. Boosting the Guessing Attack Performance on Android Lock Patterns with Smudge Attacks. In *ACM Asia Conference on Computer and Communications Security*, ASIA CCS '17, pages 313–326, Abu Dhabi, United Arab Emirates, April 2017. ACM.

[12] Geumhwan Cho, Jun Ho Huh, Junsung Cho, Seongyeol Oh, Youngbae Song, and Hyoungshick Kim. SysPal: System-Guided Pattern Locks for Android. In *IEEE Symposium on Security and Privacy*, SP '17, pages 338–356, San Jose, California, USA, May 2017. IEEE.

[13] Alexander De Luca, Marian Harbach, Emanuel von Zezschwitz, Max-Emanuel Maurer, Bernhard Ewald Slawik, Heinrich Hussmann, and Matthew Smith. Now You See Me, Now You Don't: Protecting Smartphone Authentication from Shoulder Surfers. In *ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 2937–2946, Toronto, Ontario, Canada, April 2014. ACM.

[14] Tim Forman and Adam J. Aviv. Double Patterns: A Usable Solution to Increase the Security of Android Unlock Patterns. In *Annual Conference on Computer Security Applications*, ACSAC '20, pages 219–233, Virtual Conference, December 2020. ACM.

[15] Maximilian Golla, Jan Rimkus, Adam J. Aviv, and Markus Dürmuth. Work in Progress: On the In-Accuracy and Influence of Android Pattern Strength Meters. In *Workshop on Usable Security and Privacy*, USEC '19, San Diego, California, USA, February 2019. ISOC.

[16] Maximilian Golla, Miranda Wei, Juliette Hainline, Lydia Filipe, Markus Dürmuth, Elissa Redmiles, and Blase Ur. "what was that site doing with my facebook password?": Designing password-reuse notifications. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, page 1549–1566, New York, NY, USA, 2018. Association for Computing Machinery.

[17] Marian Harbach, Emanuel von Zezschwitz, Andreas Fichtner, Alexander De Luca, and Matthew Smith. It's a Hard Lock Life: A Field Study of Smartphone (Un)Locking Behavior and Risk Perception. In *Symposium on Usable Privacy and Security*, SOUPS '14, pages 213–230, Menlo Park, California, USA, July 2014. USENIX.

[18] Taekyoung Kwon and Sarang Na. TinyLock: Affordable Defense Against Smudge Attacks on Smartphone Pattern Lock Systems. *Computers & Security*, 42(3):137–150, May 2014.

[19] Marte Løge, Markus Dürmuth, and Lillian Røstad. On User Choice for Android Unlock Patterns. In *European Workshop on Usable Security*, EuroUSEC '16, Darmstadt, Germany, July 2016. ISOC.

[20] Philipp Markert, Daniel V. Bailey, Maximilian Golla, Markus Dürmuth, and Adam J. Aviv. This PIN Can Be Easily Guessed: Analyzing the Security of Smartphone Unlock PINs. In *IEEE Symposium on Security and Privacy*, SP '20, pages 286–303, San Francisco, California, USA, May 2020. IEEE.

[21] Elissa M. Redmiles, Yasemin Acar, Sascha Fahl, and Michelle L. Mazurek. A Summary of Survey Methodology Best Practices for Security and Privacy Researchers. Technical Report CS-TR-5055, UM Computer Science Department, May 2017.

[22] Raina Samuel, Philipp Markert, Adam J. Aviv, and Iulian Neamtiu. Knock, Knock. Who's There? On the Security of LG's Knock Codes. In *Symposium on Usable Privacy and Security*, SOUPS '20, pages 37–59, Virtual Conference, August 2020. ACM.

[23] Florian Schaub, Ruben Deyhle, and Michael Weber. Password Entry Usability and Shoulder Surfing Susceptibility on Different Smartphone Platforms. In *International Conference on Mobile and Ubiquitous Multimedia*, MUM '12, pages 13:1–13:10, Ulm, Germany, December 2012. ACM.

[24] Stefan Schneegass, Frank Steimle, Andreas Bulling, Florian Alt, and Albrecht Schmidt. Smudgesafe: Geometric Image Transformations for Smudge-Resistant User Authentication. In *Conference on Ubiquitous Computing*, UbiComp '14, pages 775–786, Seattle, Washington, USA, September 2014. ACM.

[25] Youngbae Song, Geumhwan Cho, Seongyeol Oh, Hyoungshick Kim, and Jun Ho Huh. On the Effectiveness of Pattern Lock Strength Meters: Measuring the Strength of Real World Pattern Locks. In *ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2343–2352, Seoul, Republic of Korea, April 2015. ACM.

[26] Chen Sun, Yang Wang, and Jun Zheng. Dissecting Pattern Unlock: The Effect of Pattern Strength Meter on Pattern Selection. *Journal of Information Security and Applications*, 19(4–5):308–320, November 2014.

[27] Hai Tao and Carlisle Adams. Pass-Go: A Proposal to Improve the Usability of Graphical Passwords. *International Journal of Network Security*, 7(2):273–292, September 2008.

[28] Harshal Tupsamudre, Vijayanand Banahatti, Sachin Lodha, and Ketan Vyas. Pass-O: A Proposal to Improve the Security of Pattern Unlock Scheme. In *ACM Asia Conference on Computer and Communications Security*, ASIA CCS '17, pages 400–407, Abu Dhabi, United Arab Emirates, April 2017. ACM.

[29] Sebastian Uellenbeck, Markus Dürmuth, Christopher Wolf, and Thorsten Holz. Quantifying the Security of Graphical Passwords: The Case of Android Unlock Patterns. In *ACM Conference on Computer and Communications Security*, CCS '13, pages 161–172, Berlin, Germany, October 2013. ACM.

[30] Emanuel von Zezschwitz, Alexander De Luca, Philipp Janssen, and Heinrich Hussmann. Easy to Draw, but Hard to Trace?: On the Observability of Grid-based (Un)Lock Patterns. In *ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2339–2342, Seoul, Republic of Korea, April 2015. ACM.

[31] Emanuel von Zezschwitz, Malin Eiband, Daniel Buschek, Sascha Oberhuber, Alexander De Luca, Florian Alt, and Heinrich Hussmann. On Quantifying the Effective Passsword Space of Grid-Based Unlock Gestures. In *Conference on Mobile and Ubiquitous Multimedia*, MUM '16, pages 201–212, Rovaniemi, Finland, December 2016. ACM.

[32] Ding Wang, Qianchen Gu, Xinyi Huang, and Ping Wang. Understanding Human-Chosen PINs: Characteristics, Distribution and Security. In *ACM Asia Conference on Computer and Communications Security*, ASIA CCS '17, pages 372–385, Abu Dhabi, United Arab Emirates, April 2017. ACM.

[33] Guixin Ye, Zhanyong Tang, Dingyi Fang, Xiaojiang Chen, Willy Wolff, Adam J. Aviv, and Zheng Wang. A Video-based Attack for Android Pattern Lock. *ACM Transactions on Privacy and Security*, 21(4):19:1–19:31, July 2018.

Appendix

A Survey Material

Purpose of Study and Task Description

You are being asked to participate in a research study focused on the effectiveness of mobile authentication on an Android device. Androids implement pattern locks rather than traditional security parameters, for example, numeric PINs or alphanumeric passwords.

You will be asked to complete a short survey that requires you to generate a set of Android patterns under a security scenario, such as locking your device. Your eventual choices will be used in the final evaluation, as well as your responses to a set of security and usability questions.

The expected completion time of the survey is 8–10 minutes, and no more than 1 hour. You will be compensated $1.00 for your participation.

Device Usage Questions

When referring to "mobile devices" throughout this survey, consider these to include smartphones and tablet computers. Traditional laptop computers, two-in-one computers, like the Microsoft Surface, or e-readers, like the Amazon Kindle, are not considered mobile devices for the purposes of this survey.

1. How many mobile devices do you use regularly?
   ○ 0 ○ 1 ○ 2 ○ 3 ○ 4+

2. What brands of smartphone do you use for personal use? (Select all that apply)
   ☐ Apple ☐ Samsung ☐ LG ☐ Motorola ☐ Google/Pixel/Nexus ☐ Huawei ☐ ZTE ☐ Other

3. What biometric method do you use most often to unlock your primary personal smartphone?
   ○ I do not use a biometric ○ Fingerprint ○ Face ○ Iris ○ Other Biometric ○ I do not use a smartphone ○ Prefer Not to Say

   If participants indicated to use a biometric:

4a. You have indicated that you use a biometric on your smartphone. Please answer the following question related to your response. How do you unlock your primary personal smartphone when you reboot the device or if your biometric fails?
   ○ Pattern Unlock ○ 4-Digit PIN ○ 6-Digit PIN ○ PIN of other length ○ Alphanumeric Password ○ I use an unlock method not listed ○ I do not use a smartphone ○ Prefer Not to Say

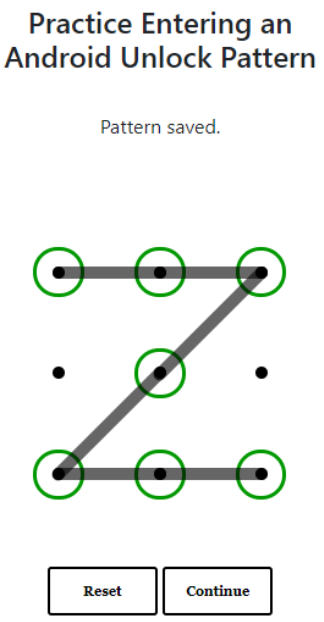


Figure 8: User interface for entering patterns.

If participants indicated not to use a biometric:

4b. You have indicated that you do not use a biometric on your smartphone. Please answer the following question related to your response. What unlock method do you use on your primary personal smartphone?
   ○ Pattern Unlock ○ 4-Digit PIN ○ 6-Digit PIN ○ PIN of other length ○ Alphanumeric Password ○ I use an unlock method not listed ○ I do not use a smartphone ○ Prefer Not to Say

What are Android Pattern Locks?

Pattern Locks are used to unlock your smartphone, like a PIN. Patterns require you to “draw” a shape that connects at least four of the contact points without lifting your finger or repeating a contact point. Displayed below is the Pattern Lock interface on a Samsung Android mobile device.

A Little Bit of Practice

On the next page, you will have a chance to practice entering an Android unlock pattern before proceeding with the rest of this survey, where we will ask you to select your own pattern that you would utilize on your primary smartphone.

Practice Entering an Android Unlock Pattern

Interface as shown in Figure 8

Instructions

For this survey, you will be asked to create an Android unlock pattern you would likely use to secure your primary smartphone.

You will need to recall this unlock pattern later in the survey, so choose something that is secure and memorable as you may use on your primary smartphone.
We ask that you DO NOT write down your patterns or use other aids to help you remember.
I understand that I should not write down my unlock pattern or use other aids to assist in the survey. ○ I understand
I understand that I will be asked to create an unlock pattern that I would use on my primary smartphone. ○ I understand

Selection

Interface as shown in Figure 8

Simple Usability Scale
Select your agreement/disagreement with the following statements. Please note that the term "system" refers to the selection of the Android unlock pattern.

5. I think that I would like to use this system frequently.
   ○ Strongly Agree ○ Agree ○ Neither Agree Nor Disagree
   ○ Disagree ○ Strongly Disagree

6. I found the system unnecessarily complex.
   ○ Strongly Agree ○ Agree ○ Neither Agree Nor Disagree
   ○ Disagree ○ Strongly Disagree

7. I thought the system was easy to use.
   ○ Strongly Agree ○ Agree ○ Neither Agree Nor Disagree
   ○ Disagree ○ Strongly Disagree

8. I think that I would need the support of a technical person to be able person to be able to use this system.
   ○ Strongly Agree ○ Agree ○ Neither Agree Nor Disagree
   ○ Disagree ○ Strongly Disagree

9. I thought there was too much inconsistency in this system.
   ○ Strongly Agree ○ Agree ○ Neither Agree Nor Disagree
   ○ Disagree ○ Strongly Disagree

10. I found the various functions in this system were well integrated.
    ○ Strongly Agree ○ Agree ○ Neither Agree Nor Disagree
    ○ Disagree ○ Strongly Disagree

11. I would imagine that most people would learn to use this system very quickly.
    ○ Strongly Agree ○ Agree ○ Neither Agree Nor Disagree
    ○ Disagree ○ Strongly Disagree

12. Select Agree as the answer to this question.
    ○ Strongly Agree ○ Agree ○ Neither Agree Nor Disagree
    ○ Disagree ○ Strongly Disagree

13. I found this system very cumbersome to use.
    ○ Strongly Agree ○ Agree ○ Neither Agree Nor Disagree
    ○ Disagree ○ Strongly Disagree

14. I felt very confident using this system.
    ○ Strongly Agree ○ Agree ○ Neither Agree Nor Disagree
    ○ Disagree ○ Strongly Disagree

15. I needed to learn a lot of things before I could get going with this system.
    ○ Strongly Agree ○ Agree ○ Neither Agree Nor Disagree
    ○ Disagree ○ Strongly Disagree

Thinking about the Android unlock pattern you just chose:

16. I feel I created an Android unlock pattern that provides adequate security for unlocking my primary smartphone.
    ○ Strongly Agree ○ Agree ○ Neither Agree Nor Disagree
    ○ Disagree ○ Strongly Disagree

17. It was difficult for me to select an Android unlock pattern that I would use to unlock my primary smartphone.
    ○ Strongly Agree ○ Agree ○ Neither Agree Nor Disagree
    ○ Disagree ○ Strongly Disagree

For participants who received a blocklist warning:
We noticed that you received the following warning while choosing your pattern:

Warning as shown in Figure 1

18a. Prior to seeing the warning above, what was your strategy for choosing your unlock pattern? [Open Text]

19a. After receiving the warning message, please describe how or if your strategy changed when choosing your unlock pattern. [Open Text]

20a. My initial strategy caused the display of this warning.
    ○ Strongly Agree ○ Agree ○ Neither Agree Nor Disagree
    ○ Disagree ○ Strongly Disagree

For participants who did not receive a blocklist warning:

18b. What was your strategy when choosing your Android unlock pattern? [Open Text]

Imagine you received the following warning message after choosing your pattern:

Warning as shown in Figure 1

19b. Please describe how or if your strategy would change as a result of the message. [Open Text]

20b. My strategy would cause this warning message to appear.
    ○ Strongly Agree ○ Agree ○ Neither Agree Nor Disagree
    ○ Disagree ○ Strongly Disagree

Recall Android Pattern
Recall the Android Unlock Pattern you created previously to secure your Primary Smartphone.

Interface as shown in Figure 8

Security Comparison
Select your agreement/disagreement with the following statements.

Questions 21-24 were shown in randomized order.

21. Unlock patterns are a secure way to unlock my primary smartphone.
    ○ Strongly Agree ○ Agree ○ Neither Agree Nor Disagree ○ Disagree ○ Strongly Disagree

22. Unlock patterns are more secure than alphanumeric passwords for unlocking my primary smartphone.
    ○ Strongly Agree ○ Agree ○ Neither Agree Nor Disagree ○ Disagree ○ Strongly Disagree

23. Unlock patterns are more secure than 4-digit PIN codes for unlocking my primary smartphone.
    ○ Strongly Agree ○ Agree ○ Neither Agree Nor Disagree ○ Disagree ○ Strongly Disagree

24. Unlock patterns are more secure than 6-digit PIN codes for unlocking my primary smartphone.
    ○ Strongly Agree ○ Agree ○ Neither Agree Nor Disagree ○ Disagree ○ Strongly Disagree

Use Unlock Pattern from Survey

25. If you were using an unlock pattern on your primary smartphone, would you use the unlock pattern you selected in this survey, or would you select a different one?
    ○ Yes, I would use the unlock pattern I created here on my primary smartphone.
    ○ No, I would not use the unlock pattern I created here and instead create a new one to use on my personal device.
    ○ Unsure, I may or may not use the unlock pattern I created here on my personal device.

26. [You have indicated that you would use / You have indicated that you are unsure if you / You have indicated that you would not use if you would use] the unlock pattern that you created in this survey on your personal mobile device. Please expand on why you [would / are unsure if you would / would not] use the unlock pattern you created here. [Open Text]

Demographics
Please enter your demographic information.

27. Select your age:
    ○ 18-24 ○ 25-29 ○ 30-34 ○ 35-39 ○ 40-44 ○ 45-49 ○ 50-54 ○ 55-59 ○ 60-64 ○ 65+ ○ Prefer Not to Say

28. With which gender do you most identify?
    ○ Female ○ Male ○ Non-Binary/Third Gender ○ Not Described Here ○ Prefer Not to Say

29. What is your dominant hand?
    ○ Left Handed ○ Right Handed ○ Ambidextrous ○ Prefer Not to Say

30. Where you live is best described as
    ○ Urban ○ Suburban ○ Rural ○ Prefer Not to Say

31. What is the shape of a red ball?
    ○ Red ○ Blue ○ Square ○ Round ○ Prefer Not to Say

32. What is the highest degree or level of school you have completed?
    ○ Some high school ○ High school ○ Some college ○ Trade, technical, or vocational training ○ Associate's Degree ○ Bachelor's Degree ○ Master's Degree ○ Professional degree ○ Doctorate ○ Prefer Not to Say

33. Which of the following best describes your educational background or job field?
    ○ I have an education in, or work in, the field of computer science, computer engineering or IT.
    ○ I do not have an education in, nor do I work in, the field of computer science, computer engineering or IT.
    ○ Prefer Not to Say

One More Thing...
Please indicate if you've honestly participated in this survey and followed instructions completely. You will not be penalized/rejected for indicating 'No' but your data may not be included in the analysis:
○ Yes ○ No

B Additional Figures and Tables

Table 5: Codebook *Pattern Select Strategy*: "Prior to seeing the warning above, what was your strategy for choosing your unlock pattern?"

Code	Frequency	Sample Quote
easy-to-remember	218	"I wanted to pick a pattern that I knew I would be able to remember."
difficult-to-guess	47	"I just started drawing something that I didn't think someone would be able to guess."
complex	31	"I tried to make a somewhat complicated pattern that I could remember."
simple	31	"Something extremely basic that I've not personally used prior to this."
unique	28	"Try to get a pattern that wasn't used a lot."
easy-to-enter	27	"I chose a pattern that would be quick and easy to use everyday."
secure	19	"I wanted something that would feel secure to lock my phone."
random	17	"I just made a random pattern that came to my mind."
many-points	13	"I tried to think of a pattern that used as many dots as possible."

* Note that each quote can be assigned multiple codes.

Table 6: Codebook *Post-Blocklist Strategy* "After receiving the warning message, please describe how or if your strategy changed when choosing your unlock pattern."

Code	Frequency	Sample Quote
complex	153	"Yes I changed it to include diagonals in a more complex manner."
difficult-to-guess	87	"Choosing a pattern that I think others would be less likely to use or guess."
different-strategy	64	"I would choose a different pattern."
easy-to-remember	55	"Didn't want to create something I'd forget quick."
long	28	"I would choose a longer one."
random	16	"I tried to think of an extremely random pattern. Something that a lot of people wouldn't select."
many-points	15	"I would use more points of the grid."
secure	13	"Make it secure."

* Note that each quote can be assigned multiple codes.

Table 7: Usage of devices and unlock methods.

	Male		Female		Other		Total	
	No.	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**
Number of Devices	624	62 %	367	36 %	15	1 %	1006	100 %
0	0	0 %	1	0 %	0	0 %	1	0 %
1	369	37 %	225	22 %	9	1 %	603	60 %
2	213	21 %	112	11 %	6	1 %	331	33 %
3	33	3 %	24	2 %	0	0 %	57	6 %
4+	9	1 %	5	0 %	0	0 %	14	1 %
Device Brand	624	62 %	367	36 %	15	1 %	1006	100 %
Apple	86	9 %	72	7 %	2	0 %	160	16 %
Samsung	231	23 %	147	15 %	6	1 %	384	38 %
LG	37	4 %	26	3 %	1	0 %	64	6 %
Motorola	39	4 %	29	3 %	0	0 %	68	7 %
Google	57	6 %	20	2 %	1	0 %	78	8 %
Huawei	9	1 %	2	0 %	0	0 %	11	1 %
ZTE	4	0 %	1	0 %	0	0 %	5	0 %
Other	161	16 %	70	7 %	5	0 %	236	23 %
None	0	0 %	0	0 %	0	0 %	0	0 %
Biometric Method	624	62 %	367	36 %	15	1 %	1006	100 %
No biometric	152	15 %	117	12 %	4	0 %	273	27 %
Fingerprint	387	38 %	183	18 %	5	0 %	575	57 %
Face	72	7 %	49	5 %	2	0 %	123	12 %
Iris	3	0 %	1	0 %	0	0 %	4	0 %
Other	7	1 %	12	1 %	0	0 %	19	2 %
No smartphone	0	0 %	0	0 %	0	0 %	0	0 %
Prefer not to say	3	0 %	5	0 %	4	0 %	12	1 %
Unlock Method	624	62 %	367	36 %	15	1 %	1006	100 %
Pattern unlock	165	16 %	95	9 %	8	1 %	268	27 %
4-Digit PIN	289	29 %	166	17 %	3	0 %	458	46 %
6-Digit PIN	100	10 %	62	6 %	3	0 %	165	16 %
Other PIN	9	1 %	8	1 %	0	0 %	17	2 %
Alphanumeric Password	33	3 %	6	1 %	0	0 %	39	4 %
Other	23	2 %	23	2 %	0	0 %	46	5 %
No smartphone	0	0 %	0	0 %	0	0 %	0	0 %
Prefer not to say	5	0 %	7	1 %	1	0 %	13	1 %

User Perceptions of the Usability and Security of Smartphones as FIDO2 Roaming Authenticators

Kentrell Owens[⋆,◇], Olabode Anise[⋆], Amanda Krauss[⋆], Blase Ur[†]
⋆ *Duo Security,* ◇ *University of Washington,* † *University of Chicago*

Abstract

The FIDO2 standard aims to replace passwords with public-key cryptography for user authentication on the web. Doing so has benefits for both usability (e.g., not needing to remember passwords) and security (e.g., eliminating phishing). Users can authenticate with FIDO2 in one of two ways. With platform authenticators, users authenticate to trusted hardware on the same device on which they are accessing a website. However, they must re-register for each website separately on each device. With roaming authenticators, such as USB security keys, they only need to register once, transferring the security key across devices. However, users might not be willing to pay for a USB security key, carry it around, or figure out how to plug it into different devices. These drawbacks have driven recent efforts to enable smartphones to serve as roaming authenticators. We conducted the first user study of FIDO2 passwordless authentication using smartphones as roaming authenticators. In a between-subjects design, 97 participants used either their smartphone as a FIDO2 roaming authenticator (via a prototype called Neo) or a password to log into a fictitious bank for two weeks. We found that participants accurately recognized Neo's strong security benefits over passwords. However, despite Neo's conceptual usability benefits, participants found Neo substantially less usable than passwords both in objective measures (e.g., timing to accomplish tasks) and in perception. Their critiques of Neo included concerns about phone availability, account recovery/backup, and setup difficulties. Our results highlight key challenges and opportunities for spurring adoption of smartphones as FIDO2 roaming authenticators.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

1 Introduction

For decades, the standard method of online authentication has involved a username and password [7, 21]. Unfortunately, password-based authentication on the web has key weaknesses. For instance, most online data breaches are caused by weak or reused passwords [48]. As a result, for decades researchers and practitioners have attempted to develop alternative authentication schemes that are more secure than passwords, yet no harder to use or deploy [7]. Consider, for example, federated identity systems like single sign-on (*SSO*). In concept, SSO eases the burden of remembering numerous passwords while also being more secure by reducing password reuse. Nonetheless, it has seen limited adoption outside organizational contexts due in part to users' privacy concerns [5, 40, 42]. Similarly, password managers have gained in popularity due to recommendations from security experts [36]. They help users generate, store, and enter unique passwords for each online account, reducing instances of password reuse or weak passwords [31]. However, adoption rates of password managers have also remained low [41].

In this line of attempts to replace passwords for web authentication, the recent FIDO2 standard [3] is a particularly promising approach that leverages public-key cryptography in place of passwords. When a user registers on a website, instead of entering a username and password, they use an *authenticator* (dedicated hardware or software following the FIDO2 specification) to generate a public-private keypair. Their client then shares the public key with the web application. FIDO2 has key benefits. In terms of usability, users no longer have to remember a password. In terms of security, users are protected from remote attacks like credential stuffing and phishing. In terms of privacy, FIDO2 does not have the centralized privacy risks of federated identity systems [13].

Major browsers Chrome, Firefox, Edge, and Safari all already support FIDO2 [33], as do a growing number of websites [47]. To use FIDO2, a user must have an authenticator, of which there are two key types. *Platform authenticators* are integrated with a broader-purpose client device and enable

authentication only on that device. For example, one can use Apple's Touch ID as a FIDO2 platform authenticator to log into websites from an iPhone or Mac laptop, or Windows Hello as a FIDO2 platform authenticator from a Windows laptop. Unfortunately, the user must re-register for a given website separately on each of their devices. In contrast, *roaming authenticators*, like USB security keys, are portable. A single roaming authenticator can be used across all of a user's devices [44]. While roaming authenticators offer usability benefits, such as enabling users to authenticate on different devices, prior work has shown users are reluctant to carry around USB security keys for authentication [10, 26]. Additionally, users may not be willing to pay for a security key.

This scenario has driven recent technical efforts to enable smartphones to be used as roaming authenticators. Over 81% of Americans own a smartphone [32], so using smartphones as roaming authenticators is likely to overcome key barriers faced by USB security keys. To this end, several proposed modifications to the FIDO2 specification are in progress, including caBLE (cloud-assisted Bluetooth Low Energy) and the closely related Network Transport [28]. Similarly, Duo Security is experimenting with a software-based mobile authenticator that we refer to as Neo. Because these implementations are recent, there has yet to be a usable security evaluation of the use of smartphones as FIDO2 roaming authenticators.

To understand user perceptions of the security and usability of Neo relative to passwords, we conducted a longitudinal user study. In a between-subjects design, participants were assigned to use either a password (termed *Password* participants) or their own smartphone as a FIDO2 roaming authenticator via the Neo prototype (termed *Neo* participants) to log into a fictitious bank from their own computer daily for two weeks. A total of 97 participants completed the full protocol and all daily tasks. We asked a series of research questions:

- **RQ 1**: Neo involves non-trivial setup relative to passwords. How difficult do users find Neo's initial setup?

By both objective and subjective measures, participants found Neo's setup process difficult. More than half of *Neo* participants dropped out of the study before completing the setup process, whereas under 10% of *Password* participants did so. Even among those who did complete setup, it took the median *Neo* participant over fifteen minutes to configure the software. While some of this difficulty was due to Neo being a research prototype, other aspects were inherent in using smartphones as roaming authenticators. Even beyond one-time setup costs, the recurring steps in account creation took longer for *Neo* participants than for *Password* participants.

Neo participants also perceived the setup process as less usable than *Password* participants. In particular, *Neo* participants rated the setup process 20 points lower on the 100-point system usability scale (*SUS*) than *Password* participants.

- **RQ 2**: In daily authentication, how does the usability of Neo compare to passwords (after Neo has been set up)?

Passwords can be forgotten or mistyped, whereas Neo simply requires access to a smartphone. Thus, we expected daily authentication to be easier for Neo than for passwords. We found the opposite, however. *Neo* participants were more likely than *Password* participants to be *unsuccessful* at logging in, typically because they could not authenticate to their phone (e.g., with their fingerprint sensor) or their phone seemed not to receive the push notifications that are part of the protocol. Unsurprisingly, then, *Neo* participants were less likely to rate daily sign-ins as easy than *Password* participants.

- **RQ 3**: Overall, how do users perceive the security and usability of Neo relative to passwords? Are they correct?

Overall, participants perceived Neo as both secure and usable. Notably, participants correctly perceived Neo as *more secure* than passwords. However, they also perceived Neo as *less usable* than passwords even beyond their direct experiences with setup and authentication. Nonetheless, over half of the participants who used Neo reported being "likely" or "very likely" to use Neo over passwords for five of the six account types (all except banking) that we asked about.

- **RQ 4**: Collectively, what are the barriers to user adoption of smartphones as FIDO2 roaming authenticators?

Neo participants frequently expressed concerns about not having their phone available or accessible when they hoped to log in. Notably, one-third of participants reported misplacing their phone at least once a day. They also worried about account recovery and losing access to their account, whereas they could simply write their password down somewhere safe. As a result, many participants expressed reluctance to adopt Neo for their own accounts even after using it.

- **RQ 5**: Does a user's prior experience with two-factor authentication (*2FA*) influence their perceptions of Neo?

We found that *Neo* participants who had prior 2FA experience rated its usability more highly (in terms of SUS score) than those who had never used 2FA.

Collectively, our work contributes the first user-centered understanding of smartphones as FIDO2 roaming authenticators. We uncovered a number of usability drawbacks, both in actuality and in perception, that will likely hamper the adoption of systems like Neo even as they move from research prototypes to being directly integrated with browsers. We thus highlight key challenges and opportunities for spurring adoption of smartphones as FIDO2 roaming authenticators.

Our paper proceeds as follows. We first detail how FIDO2 works (Section 2) and then present our user study's methodology (Section 3). Next, we present our study's results (Section 4). We then discuss these results (Section 5), compare them with related work (Section 6), discuss both limitations and future work (Section 7), and finally conclude (Section 8).

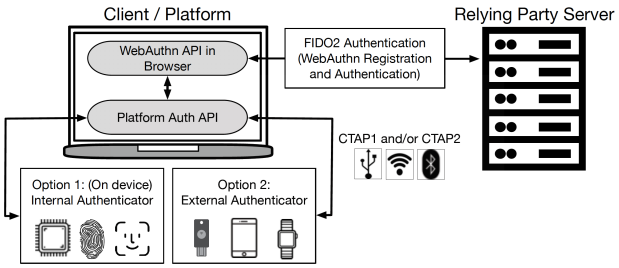


Figure 1: FIDO2 authentication with WebAuthn and CTAP2. This diagram is taken from Lyastani et al. [26].

2 Background

In this section, we further detail FIDO2 and its constituent protocols on a technical level. We particularly focus on efforts to support smartphones as roaming authenticators. Figure 1 summarizes the FIDO2 authentication process.

2.1 FIDO2: WebAuthn and CTAP2

The FIDO2 standard includes two key protocols. The Web Authentication API (*WebAuthn*) is a standard jointly developed by the FIDO Alliance and the W3C [33]. The WebAuthn API enables web applications (termed *relying parties*) to leverage public-key cryptography to authenticate users. Instead of a password, a unique public/private key pair is generated for each website registration using an authenticator. The private key is stored on the user's authenticator. The public key, along with a randomly generated credential ID, is stored on the web application's server. Credentials are scoped to the web application through the use of a relying party identifier that identifies the server. The user can then authenticate to that web application by interacting with their authenticator.

The other half of FIDO2 is *CTAP2*, a protocol being developed by the FIDO alliance. It is used when a relying party is interacting with a roaming authenticator [3], such as mobile devices like smartphones. The two salient parts of the protocol are the Authenticator API and the transport-specific bindings, referred to as *transports*, that can be used. The Authenticator API details how an authenticator should interact with a relying party when making a credential (i.e., public/private key pair) and creating assertions that provide proof of an authentication and a user's consent. The protocol defines how each of these operations should take place given the capabilities of the authenticator. The transports are how messages are conveyed from the host to a roaming authenticator. Currently, the modes that are supported are USB, NFC, and Bluetooth. The next section details implementations using these transports.

2.2 Mobile Roaming Authenticator Efforts

Next, we summarize three recent efforts that enable mobile devices to be used as FIDO2 roaming authenticators.

simFIDO is an implementation of FIDO2 by Chakraborty et al. [8] that uses a SIM-card-based Trusted Platform Module (TPM) called simTPM [9] to allow Android devices to serve as hardware authenticators. The authors introduced a new Android system service called *External FIDO Request Receiver Service* (XFRR) that forwards CTAP commands to the simTPM. Unlike typical implementations where credentials are bound to a particular device and cannot be removed, a SIM card (the authenticator) can be moved across devices.

caBLE (Cloud-Assisted BLE) is a proposal by Google that would extend CTAP2. It attempts to overcome some of the disadvantages of system BLE pairings, such as client-implemented preference syncing. The caBLE proposal allows mobile devices to serve as a roaming mobile authenticator by establishing a secure channel to pass CTAP2 messages between the authenticator and the client (e.g., the Chrome browser) [28]. The latest version of this proposal, caBLEv2, permits both temporary and permanent pairings between devices. The latter is appropriate for a personal device.

Neo is a prototype developed by Duo Security that allows mobile devices to serve as roaming authenticators. To use Neo, the user first pairs their mobile device with a Chrome browser with the aid of a mobile application and Chrome extension. The pairing process between the mobile device and the client takes place through a QR code generated by the extension. The QR code contains a shared secret. After the successful pairing, the client communicates with the mobile device through proxying of the WebAuthn API actions via the Chrome Extension to an intermediary server. Whenever the user attempts to authenticate on a website, they will receive a push notification to their smartphone that they can accept or reject after unlocking their device (if their phone is locked). Ongoing work aims to add an HTTPS-based transport (Network Transport) to the list of CTAP2 transports [2, 28]. With the addition of Network Transport to the CTAP2 specification, the Chrome extension would no longer be necessary during assertion or pairing for Neo or similar efforts; CTAP2 authenticators could communicate with the client directly.

There is debate about when user *presence*, versus user *verification*, should be required for authentication. Yubico recommends *presence* for 2FA and *verification* for passwordless authentication (like Neo) [46]. Since simple possession of a device is insufficient for authentication, we predict that similar schemes will (like Neo) require users to unlock their device before responding to an authentication push request; not doing so facilitates many attacks. Account sharing is easy with FIDO2 — simply register multiple phones with one account. Shared phones would be a security risk and potentially not possible if biometric user verification is required.

3 Methodology of Our User Study

To understand users' initial perceptions of the security and usability of using a smartphone as a FIDO2 roaming authen-

ticator, as well as how those perceptions might change after extended use, we conducted a longitudinal, between-subjects study. We compared the relative usability of passwords and Neo using both qualitative and quantitative methods. This study was conducted between May 2020 and July 2020.

3.1 Recruitment

We recruited participants on Amazon Mechanical Turk (*MTurk*), an online crowdsourcing platform that has been frequently used in usability studies. Redmiles et al. found that MTurk users are often more diverse in terms of age, income, education level, and geography than traditional social science pools [34]. Because of the requirements for using Neo (Chrome extension and Neo app), we required participants to have an Android mobile phone, access to a computer, and Google Chrome installed on that computer. Participants had to complete a screening survey verifying that they met the requirements for participating in the study, including 1) having an Android mobile phone running Android version 9+, 2) being located in the US, 3) using Google Chrome, and 4) having a fingerprint scanner on their phone. We used a free web service to validate the location of participants, following techniques outlined by Kennedy et al. [22]. Participants also had to have a 95% approval rating on MTurk.

3.2 Study Design

Eligible participants (n=247) were randomly split into two groups, with each group assigned one authentication method (*Password* or *Neo*). We informed participants that they were participating in a study about online authentication and that they would perform a series of ten tasks on a fictitious banking application over the course of two weeks. The banking application we used was a fork of the one used by Reese et al. [37] in a prior study [38], modified to support FIDO2.

We then instructed each group on how to register for an account with our web application using their assigned authentication method. Participants assigned to the *Password* condition were instructed to choose a username and password. We required that the password chosen contain at least 8 characters, without any further restrictions. Similar to Lyastani et al. [26], we chose this password-composition policy, which is the simplest NIST-recommended password policy, to avoid skewing usability perceptions with a potentially frustrating password-composition policy [43]. Choosing a more complex password-composition policy may have led to different perceptions of both the security and usability of passwords. Furthermore, to replicate participants' current approach to passwords, we neither encouraged nor prohibited the use of a password manager. Participants assigned to the *Neo* condition were given instructions on how to install the mobile application and Chrome extension needed for the Neo prototype to work, how to complete the pairing process between the mobile phone and the Chrome Extension, and how to register for an account on the banking website.

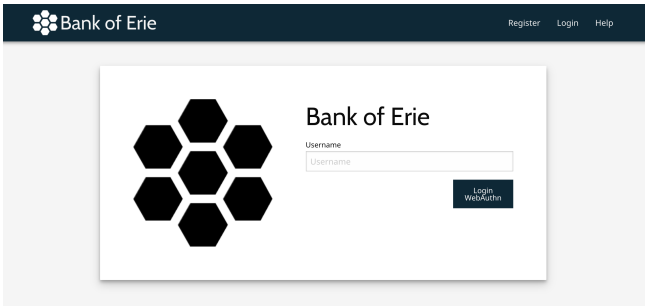


Figure 2: A screen shot of the simulated banking application to which participants authenticated throughout the study.

After successfully registering for their account and logging into the web application, participants were then instructed to complete one of the ten required tasks by using their assigned method to log into the banking application. Participants then completed the System Usability Scale (*SUS*) [16], evaluating the usability of setting up their assigned authentication method. In addition, they answered questions concerning their experience with setup and provided demographic information.

Over the two-week period, participants were sent a daily reminder to complete one of the ten required tasks by authenticating to the banking application. Full participation in the study required completing these ten tasks within 14 days. At the conclusion of the study, participants in both groups completed an exit survey where they completed another SUS questionnaire and answered open-ended questions about their authentication experience during the two weeks.

3.3 Attrition

During the development of the study protocol, we expected participant attrition in both the *Password* and *Neo* groups because of the longitudinal nature of the study. For participants in *Neo*, specifically, we hypothesized that there were steps that could prove to be challenging, causing additional attrition. Two such steps were app installation and fingerprint enrollment. For participants to install the Neo prototype on their Android, they would have to sideload the application. Depending on their version of Android, doing so necessitated enabling a setting to install unknown applications. While we detailed this process in our onboarding instructions, it is possible that participants did not feel comfortable enabling the setting. To use Neo, participants would also have to register a fingerprint if they had not already done so. This fingerprint was used to authenticate to their phone, thus instructing their phone to use their private key to authenticate to the banking application. Given common misconceptions about biometrics [6, 25], we anticipated that some participants assigned to *Neo* would not feel comfortable enrolling a fingerprint. Not

enrolling a fingerprint would prevent them from completing onboarding and participating in the remainder of the study. Sections 4.2–4.3 detail the attrition rates observed in practice.

3.4 Data Collected

We hypothesized that one of the barriers to the adoption of FIDO2 in general, and specifically a roaming mobile authenticator, would be a poor setup experience. For *Neo* participants, there are several steps where a participant could get stuck or have difficulty, such as installing the Chrome extension or pairing the Chrome extension and the mobile application. To understand which steps proved to be the most problematic, we collected detailed timing data on the following parts of the setup process for *Neo*: 1) downloading the Chrome extension, 2) enrolling a fingerprint, 3) downloading the mobile application, 4) pairing the Chrome extension and mobile application, and 5) registering a credential on the experiment platform. Since password-based authentication is something that *Password* participants would be familiar with, we only collected timing data for the account creation step. We attempted to capture each participant's initial impressions of the usability of their assigned authentication method through both the SUS questionnaire and a series of open-ended questions.

During the longitudinal portion of the study, we had three goals. First, we wanted to understand how long and error-prone the login experience was over time. To do so, we collected data on how long it took participants to authenticate during each of the ten sessions, recording failed authentication attempts (whether due to timeouts/cancellations in the *Neo* condition or incorrect password entries in the *Password* condition). Second, we wanted to understand how participants felt in the moment after each authentication. We accomplished this by implementing a diary-style Likert item where we asked participants to rate their agreement ("strongly disagree" through "strongly agree') that "logging into this application is easy." Lastly, we wanted to understand how participants' opinions of their assigned authentication method changed over time. Thus, we again asked participants to complete an SUS questionnaire and answer relevant open-ended questions.

3.5 Data Analysis Methods

We conducted both qualitative and quantitative data analyses. For free-response data, we conducted qualitative content analysis. In particular, two researchers independently read through the full survey data, each making a broad list of topics participants raised. They discussed the list and jointly created a code book combining topics under closely related themes. They iterated on this code book and reached consensus on the codes they would use. Using these codes, they both independently coded one-third of the responses. Cohen's κ, a measure of inter-coder agreement [11], was 0.85 between the two researchers. Fleiss et al. [19] consider values of κ over 0.75 as excellent agreement. If the two researchers disagreed on a code, they subsequently discussed the disagreement and reached consensus. After observing this acceptable value of κ, one researcher independently coded the remaining responses.

Many of our quantitative analyses were comparisons between the *Neo* and *Password* groups, such as in timing or SUS scores. Because most of this data was not normally distributed, we typically used the Mann-Whitney U test (*MWU*, also known as the unpaired two-samples Wilcoxon test). We also sought to understand how participants' success at authenticating, time required to authenticate, and Likert-scale responses both changed over time (across the ten authentication sessions) and varied between the *Neo* and *Password* groups. Because this data was not independent, we built mixed-effects regression models with the participant as a random effect. Based on the outcome data types of these three longitudinal models, we respectively built mixed-effects logistic, linear, and ordinal regression models. For all statistics, $\alpha = .05$.

3.6 Ethics

While Duo Security is not an academic institution and does not have a formal IRB, the study protocol and mechanisms used to conduct the study underwent an internal privacy review. As part of the screening process, participants had to consent to their data potentially being published externally. To conduct the study, we needed to store usernames and hashed versions of participants' passwords, but that data was discarded at the conclusion of the study. Participants were eligible for up to $30 in compensation based on whether they completed the survey that followed the setup process, the ten daily authentication tasks, and the exit survey. We chose that number to compensate participants at roughly $15/hr.

3.7 Pilot Study

Prior to conducting the study on MTurk, we conducted an internal, eight-person pilot to uncover potential problems in our study setup (e.g., survey questions, mobile app user interface, time allocated, bugs in the experiment platform), as well as to identify additional questions to ask. At the conclusion of the pilot, we corrected bugs in the code for the experimental platform. We also made our setup instructions more clear.

4 Results

In this section, we describe our participants and then report our results, grouped chronologically and by research question.

4.1 Participants and Their Demographics

Overall, 97 participants completed all parts of the protocol: onboarding, initial survey, ten daily tasks, and the exit survey. While the initial assignment to groups was randomized

Table 1: The demographics of the 97 participants who completed all surveys and parts of the full longitudinal protocol.

	Password	*Neo*
Gender		
Female	31	10
Male	28	17
No Answer	7	4
Age		
18–24 years old	2	5
25–34 years old	26	12
35–44 years old	26	7
45–54 years old	8	4
55–64 years old	1	2
No Answer	3	1
Race		
American Indian or Alaska Native	0	1
Asian	5	5
Asian, White	1	0
Black or African American	2	2
Black or African American, Hispanic	1	0
Hispanic	6	1
Hispanic, White	3	1
White	45	20
No Answer	3	1
Education		
High School Diploma/GED	3	3
Some College But No Degree	13	6
Associate's Degree	10	5
Bachelor's Degree	24	12
Professional Degree	13	4
No Answer	3	1
CS Background		
Yes	4	8
No	59	22
No Answer	3	1
TOTAL	66	31

and approximately equal, 66 participants in *Password* and 31 participants in *Neo* completed all parts of the protocol. We discuss this unequal attrition between groups further in Sections 4.2–4.3. All analyses other than those of participant attrition report on these 97 participants.

Table 1 summarizes participants' demographics. Participants in both conditions were more educated than the broader United States population, with 59% of *Password* participants and 53% of *Neo* participants having attained a bachelor's degree or higher. We also asked which of five 2FA methods (SMS, TOTP, pre-generated codes, push-notification based, and security keys) participants had used before; we included example images of these different methods to make them easily identifiable. Among participants, 94% had used SMS, 54% had used TOTP, 45% had used push notifications, 26% had used pre-generated codes, and 3% had used security keys. The proportions of each were similar between conditions.

Table 2: Summary of the time (in seconds) to set up Neo.

Step	10th %-ile	Median	Mean	90th %-ile
Install Chrome Extension	30.4	56.7	283.5	155.4
Enable Fingerprint	5.5	12.1	73.8	162.5
Install Phone Application	48.1	96.6	220.9	414.1
Pair Device	52.7	148.2	218.0	505.0
Create Account	17.9	105.3	139.4	233.9
Total Time	432.9	1000.1	1488.9	2316.3

4.2 Initial Setup (RQ 1)

As mentioned in Section 3.4, we measured the time it took for *Neo* participants to complete each step in the setup process. The timing data was collected through Qualtrics as participants progressed through the setup guide. The median time to complete setup for *Neo* was 1,000.1 seconds (16 minutes and 40.1 seconds). The step with the highest median time was pairing the participant's mobile device with the browser, which took 148.2 seconds. The step that took the least time was enabling fingerprint authentication for participants who did not already have it enabled. The majority (23/31) of participants in *Neo* reported that they already had their fingerprints enrolled on their smartphone prior to beginning the study. Additional timing results for *Neo* can be found in Table 2.

The only step in the setup process that *Neo* and *Password* both shared was account creation. The median time for account creation was 74.4 seconds for *Password* and 105.3 seconds for *Neo*, though this difference was not statistically significant (MWU, $U = 1269$, $p = .142$). In addition to the timing data, we analyzed the SUS scores that participants submitted after they completed setup. The median SUS score for *Password* was 88.6, while for *Neo* it was 66.6. This difference was significant (Mood's median test, $p < .001$).

Of the 31 *Neo* participants who completed the setup process, 11 nonetheless described challenges they encountered. They described it as too complex, particularly the process of downloading the mobile app and pairing the phone with the browser. P5 managed to get it working, but expressed frustration with the process: *"I never really understood exactly what I was doing or what was required when logging in. I figured out the steps to make it work but don't understand the meaning or process."* Five participants said the installation and setup process should be simpler. One participant suggested adding video instructions to the text ones provided, while another suggested that the additional app download and the extension should be eliminated entirely, if possible. Those steps could indeed be eliminated if Neo were supported natively in future web browsers, though usability challenges would remain.

An important usability finding is that we observed a high rate of participant attrition and drop-out, particularly among *Neo* participants during the setup phase. To control access to the study, we utilized MTurk qualifications. After assign-

ing groups randomly among eligible participants from the screening survey, there were 123 participants in the *Password* condition and 115 participants in the *Neo* condition. At the conclusion of the setup phase, only 45% (52) of the participants assigned to *Neo* remained. Comparatively, 91% (112) of the participants assigned to *Password* remained. This difference in attrition during onboarding across conditions was significant ($\chi^2(1) = 64.596, p < .001$).

At each stage of the setup process for *Neo*, we saw participants leave the study. The two steps that resulted in the worst attrition were installing the application (16 participants dropped out) and creating an account (15 participants dropped out). As mentioned in Section 3.3, we posited that installing the application would cause problems. However, we did not predict that account creation would trail so closely. It is possible that participants did not have a compatible device to complete the credential creation process, or they could have reached a threshold of frustration with the entire onboarding process. In Section 4.1, we detailed that our final sample for *Neo* consists of 31 participants who completed all phases of the study. The remaining attrition occurred longitudinally.

4.3 Daily Authentications (RQ 2, RQ 5)

In the longitudinal phase, participants authenticated ten times over 14 days. We measured the additional participant attrition, errors logging in, the time authentication took, and participants' perceptions of the usability of logging in.

Attrition: During the longitudinal portion of the study, participants left the study at similar rates across the *Password* and *Neo* groups ($\chi^2(1) < .001, p = 1.000$). Among participants who successfully completed the setup process, 60% of *Neo* participants and 59% of *Password* participants completed all remaining parts of the full protocol.

Authentication Errors: Participants attempted to log into the banking application ten times in 14 days, and we recorded each time whether they successfully authenticated using their assigned mechanism. Although we provided no training for the *Neo* condition, Figure 3's jump in authentication success rate from Day 1 to Day 2 demonstrates quick learning. Across authentication attempts for all days, 98% of attempted *Password* authentications were successful, while 87% of attempted *Neo* authentications were successful.

To quantify how authentication failures varied across groups and changed over time, we created a mixed-effects logistic regression model. Authentication success was the dependent variable, while the assigned group, the day in the study, and the interaction of those terms were the independent variables. As this data is not independent, the participant was modeled as a random effect. Table 3 presents our model. As suggested above, we found that *Neo* participants were less likely than *Password* participants to authenticate successfully

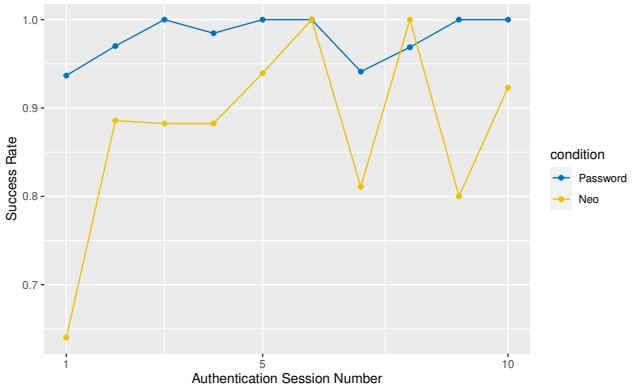


Figure 3: The success rate of authentication attempts over the ten authentication sessions, split by condition.

($OR = 0.222$, $p = .028$). We observed a marginally significant effect in that a participant was more likely to authenticate successfully the further the participant was in the fourteen-day longitudinal protocol ($OR = 1.179$, $p = .056$). Figure 3 shows that the lowest rate of successful *Password* authentications (94%) occurred during the first authentication session of the study. Comparatively, the authentication success rate was 64% for *Neo*. We defined an authentication session as all authentication attempts that occurred in a 10-minute span.

Fingerprint scans were one cause of errors for *Neo* participants. Participants noted that the reliability of a fingerprint scanner varies, and they may not work in certain scenarios (e.g., when one's finger is wet). A few participants mentioned that they did not use biometrics on their phone before this study, and others mentioned issues with their fingerprint scanners during the study. P11 described how they *"had to add extra finger scans into [their] phone in order to get it to work better."* Some participants brought up reliability issues with the Neo platform prototype itself, saying that they sometimes had to try multiple times to receive push notifications.

Timing Data: For each authentication attempt, we logged when the user landed on the login page and when they completed authentication (pressing submit or approving the push). The average times to authenticate for *Neo* and *Password* were 20.9 seconds and 8.1 seconds, respectively. In our mixed-effects linear regression model (Table 4), we found that *Neo* participants took significantly longer to authenticate than *Password* participants ($\beta = 13.708$, $p < .001$). We also observed a marginally significant result that authentication took slightly less time as the study progressed ($\beta = -0.217$, $p = .083$).

Ease of Authentication: After each authentication, participants responded on a Likert scale ("strongly disagree" to "strongly agree') to the statement "logging in to this application is easy." We again built a model, this time a mixed-effects ordinal regression model (Table 5). We found that *Neo* par-

Factor	Baseline / (Type)	Odds Ratio	95% CI	σ	z	p
Group: *Neo*	Password	0.222	[0.058, 0.850]	0.685	-2.198	**.028**
Day in Study	(Continuous variable)	1.179	[0.996, 1.397]	0.086	1.914	.056
Group: *Neo* * Days in Study	(Interaction effect)	0.909	[0.745, 1.108]	0.101	-0.947	.344

Table 3: A mixed-effects logistic regression model of participants' success (1) or failure (0) logging in on each day of the longitudinal study. The independent variables (*IVs*) were the participant's assigned group and how many days into the longitudinal study they were, as well as the interaction between the two. We report the odds ratio and 95% confidence interval of the odds ratio (*95% CI*). We also note the baseline (for categorical predictors) or the data type of the IV, as applicable.

Factor	Baseline / (Type)	β	95% CI	SE	DF	t	p
Group: *Neo*	Password	13.708	[8.574, 18.841]	2.627	168.773	5.219	**<.001**
Day in Study	(Continuous variable)	-0.217	[-0.462, 0.028]	0.125	910.706	-1.734	.083
Group: *Neo* * Days in Study	(Interaction effect)	-0.219	[-0.673, 0.235]	0.232	907.666	-0.943	.346

Table 4: A mixed-effects linear regression model of the time it took participants to authenticate on each day of the longitudinal study. The IVs and terminology are the same as in Table 3.

ticipants consistently had lower agreement that logging in was easy compared to *Password* participants ($OR = 0.012$, $p < .001$). As the study progressed, participants had higher agreement that logging in was easy ($OR = 1.134$, $p = .002$), even more so in the *Neo* group ($OR = 1.140$, $p = .022$). When examining the data over time for *Neo*, we found that the lowest percentage of responses agreeing or strongly agreeing that logging in was easy (73%) occurred on the first day of the study. Across the entire study, 90% of the *Neo* responses agreed or strongly agreed that logging in was easy. This number was 99% for *Password*.

Usability: At the study's conclusion, we asked participants to complete an additional SUS questionnaire to understand if their perceptions of the usability of Neo had changed over the course of the study. Figure 4 summarizes the results from the exit SUS. The average *Neo* SUS score was 81.3 in the exit survey, compared to 66.6 in the initial survey. When transforming those scores using the adjective scale from Bangor et al. [4], Neo received an "OK" rating in the initial survey and a "Good" rating in the exit survey. Comparatively, passwords received an "Excellent" rating in both surveys, with average SUS scores of 88.6 and 90.4 for the initial and exit surveys, respectively. Exit survey SUS scores for *Password* were higher than for *Neo* (MWU, $U = 1393.5$, $p = .002$).

SUS scores for *Neo* were also higher in the exit survey than in the initial survey (Paired Wilcoxon, $W = 636.5$, $p = .003$). As P12 said, *"While [Neo] may seem unfamiliar, it is quick to set up and easy to learn, and it makes getting into your account quick and straightforward."* Although *Neo* received lower SUS scores than passwords, 23 of 31 *Neo* participants nonetheless described Neo as "simple," "easy-to-use," or "straightforward" in free-response data. We also found that participants with prior experience with push notifications for 2FA found Neo more usable (MWU, $U = -1.965$, $p = .050$).

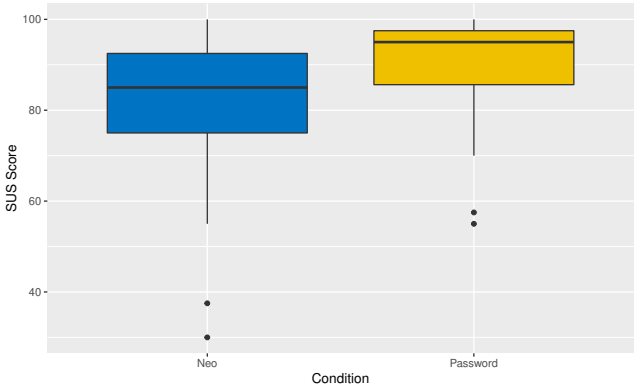


Figure 4: SUS scores from exit survey by condition

4.4 General impressions of *Neo* (RQ 3, RQ 4)

Security: All *Neo* participants expressed a belief that Neo was secure or trustworthy because of either its requirement for physically possessing one's phone or the use of biometric fingerprint scans.

> P31: *"I definitely think this authentication method makes online accounts safer. The fact that it sends a notification to your phone and requires your fingerprint makes me feel that my account is safe and secure because only I can authenticate the logins."*

P27 even mentioned that they believed Neo protected them from *"the problem of SIM card hacking."*

Availability: Participants mentioned a number of concerns about using Neo for authentication. The most common concern was phone availability (18 participants). Some participants described needing to have your phone physically nearby as annoying: *"It's a minor annoyance to try to log in when*

Factor	Baseline / (Type)	Odds Ratio	95% CI	σ	z	p
Group: *Neo*	Password	0.012	[0.002, 0.064]	0.832	-5.270	**<.001**
Day in Study	(Continuous variable)	1.134	[1.045, 1.230]	0.042	3.024	**.002**
Group: *Neo* * Days in Study	(Interaction effect)	1.140	[1.019, 1.275]	0.057	2.298	**.022**

Table 5: A mixed-effects ordinal regression model of participants' agreement (5 = "strongly agree"; 1 = "strongly disagree") that "logging in to this application is easy" on each day of the study. The IVs and terminology are the same as in Table 3.

your phone is across the room or in another room charging" (P4). Participants also described other availability challenges, such as a phone running out of battery, not having internet access, or being broken. For example, P24 wrote, *"If the phone breaks or is forgotten somewhere (I know this is probably uncommon), I didn't really see an alternative way to log in or secure your account."* When asked at the end of the study how frequently they were generally unable to access their mobile device when they needed it, 10 participants said "once a day," while the rest said at most "once a week." Notably, 10 other participants said that this "almost never" happened.

Related to availability challenges, participants raised concerns about account recovery or a backup authentication method. They pointed out how Neo lacked an obvious recovery/backup method, and this could cause them to be locked out of their accounts. For example, P32 wrote, *"If I can't authenticate with my phone and there is not a backup login procedure, then I can't login to my account."*

Privacy: Three participants were concerned about the privacy implications of using Neo. Regarding fingerprints, P13 wrote, *"[Neo] requires a thumbprint on the phone currently to use it, so people concerned about the privacy of that cannot use it."* One participant mentioned that they feel like Neo gives the banking institution too much information, while another said they would not be comfortable setting up Neo on someone else's client while away from their computer.

Deployment: Fifteen participants offered concrete suggestions for improving Neo. Participants mentioned wanting alternatives to using a fingerprint for locally authenticating to their smartphone (as part of the process of the phone serving as a roaming authenticator). They suggested facial recognition, a PIN, or behavioral authentication. P15 wrote, *"This is probably far-fetched but maybe in the future . . . it just knows you are the one holding the phone. Instead of giving me a popup to select an action, it simply registers your fingerprint when holding the phone and logs you in."* Participants voiced the need for account recovery/backup methods to be available. Finally, some participants commented that the UI was plain and should be improved.

Adoption: Eight participants mentioned (unprompted) that they would use Neo if it were widely available for authentication. One additional participant said they would use it for 2FA, but not as their primary form of authentication. P24 wrote, *"For relatively unimportant account (like dating or streaming services), this is already enough for me to use it as long as I feel like Neo is a trustworthy and secure company."* Four participants explicitly stated that they believe the benefits of Neo outweigh the additional effort it requires. For example, P27 wrote, *"It is a little more 'difficult' than just entering a password, since it involves another step (grabbing your phone and opening up the authentication app), but the added security makes it worth it."*

When asked how likely ("very likely" to "not likely") they would be to use Neo over passwords for six different account types (dating services, streaming services, social media, health care services, banking, and email), over half of *Neo* participants said they were "likely" or "very likely" to use Neo over passwords for all account types except for banking. Streaming services ranked the highest with 61%. Banking ranked the lowest, with less than half (39%) of *Neo* participants being likely to use Neo over a password.

4.5 Comparisons with Alternatives (RQ 4)

Some *Neo* participants made comparisons between Neo and other authentication schemes they had used. Thirteen participants described benefits they perceived Neo as having relative to passwords. These benefits (in order of decreasing prevalence) included not having to remember/store passwords, the security benefits of using biometrics (instead of a password that might be cracked), and ease of use. For example, P13 wrote, *"It's also very easy to use because you just have to use your thumbprint to verify that it's you rather than taking the time to type out a password and guessing which password you used for which account."* Individual participants also mentioned that they found Neo easier to use than password managers or email/SMS PINs. Conversely, one participant described Neo as frustrating relative to alternative schemes:

> P11: *"I didn't really like it. I thought it was a bunch of extra unnecessary steps just to log in and do some simple tasks. It got easier to use, but was still clunky and I really didn't like it . . . There are easier and better ways to do authentication that aren't as frustrating or unnecessary."*

Some participants described Neo as being similar to 2FA. For example, P18 wrote, *"It provides authentication like 2FA. I feel it makes things somewhat safer."*

Password participants confirmed findings from prior work on passwords [7], including that they found passwords simple, familiar, and easy to use. Some participants specifically called out how learnable passwords are, even for people with little technical expertise: *"The authentication was easy to use and not too complex . . . It is easy to use for those with very little computer knowledge or skills"* (P49). A few *Password* participants mentioned that they liked that they did not need a second device to authenticate. Others described the historical resiliency of passwords as a sign of its strength as an authentication scheme: *"It's been proven to be quite secure (when done properly) over decades of use"* (P95) and *"No one has yet come up with something worth the trade-offs"* (P91).

When asked about disadvantages of passwords, *Password* participants overwhelmingly mentioned security. Their concerns regarding password security included weak passwords, others learning one's passwords, password reuse, and the risks of browsers' auto-fill login if someone gains access to their devices. Several participants said that our fictitious banking application's password policy should have been stronger to ensure they created secure passwords. One participant also discussed the lack of a CAPTCHA on the login page, mentioning how bots could hack accounts. Several participants raised the lack of multi-factor authentication (*MFA*) as a weakness of our implementation of password authentication:

> P40: *"Nowadays, it feels a little vulnerable for something like banking not to require a two-step validation process using a texted or emailed validation code. . . . If there was no second step to verify the user generally, I might be a little concerned."*

Participants suggested different forms of MFA that could improve passwords, including an email/SMS code, security questions, physical presence, and biometrics (facial recognition, fingerprints). As P76 wrote, *"I still think two-factor authentication can protect the safety of online accounts better on top of the traditional password authentication method. [Add] face recognition, SMS/email/app authentication, physical authentication assure the users that they're more protected."*

5 Discussion

In this section, we discuss our results' implications for efforts to spur adoption of smartphones as roaming authenticators.

5.1 Separating Setup and Day-to-Day Use

Reynolds et al. recommended that researchers study setup and day-to-day authentication separately when evaluating authentication schemes so that problems during the setup phase would not impact participants' perceptions of usability for day-to-day use [39]. In our initial study plan, we intended to help participants set up Neo in-person. However, the COVID-19 pandemic shifted our study online and changed plans such that participants had to set it up themselves. Given the improvement in SUS scores for both conditions between our initial and final surveys, we believe that the overall experience for participants in both condition was impacted by their setup experience. This is specifically evident in the significant portion of participants who dropped out of the study before completing the setup process. In our analysis of the longitudinal data and initial SUS scores, we found that *Neo* participants had comparatively worse experiences authenticating at the beginning of the study. However, as the study progressed, participants found authenticating somewhat easier, authenticated somewhat faster, and made somewhat fewer authentication errors. Moreover, *Neo* participants found the scheme more usable at the conclusion of the study than at the beginning. Separating setup and daily use also enabled us to disentangle perceptions of Neo from general perceptions of using smartphones as roaming authenticators. As different technical approaches develop for using smartphones as roaming authenticators, their implementations may have significantly different setup processes. To better understand perceptions of smartphones as roaming authenticators after continued use, researchers will need to evaluate the day-to-day use of FIDO2 implementations like Neo separately from setup.

5.2 Security vs. Usability

Timing data showed that *Password* participants authenticated more quickly than *Neo* participants at every point in the study. We attribute the authentication speed for *Password* participants to both familiarity and the use of auto-fill capabilities by password managers and browsers. 25% of *Password* participants reported that they used a password manager, browser, or other tool to generate new passwords. These timing results are similar to the findings of Farke et al. [18]. Neo's consistent underperformance relative to passwords raises the question of whether highlighting an authentication method's security benefits is enough to encourage adoption.

Unlike in prior work on security keys, in which participants did not fully understand the potential benefits of security keys [14, 26], participants in our study reported that Neo was substantially more secure than passwords, yet found passwords more usable. Nonetheless, the majority of participants who used Neo during the study reported being likely to use Neo over passwords for all account types we asked about other than banking accounts. It is possible that users of password managers already receive FIDO2's best non-security related attributes (e.g., memorylessness, decreased cognitive load during registration). To counter similar arguments and spur adoption, implementers will need to underscore the flaws of passwords, such as the threat of phishing, credential stuffing, and data breaches, highlighting how FIDO2 avoids them.

5.3 Availability/Account Recovery

The most common concern for *Neo* participants was phone availability. For people to feel comfortable adopting smartphones as roaming authenticators, system designers must solve availability issues that arise from using a smartphone to authenticate. That is, if the user's smartphone is their only authenticator and their smartphone is inaccessible for any of the reasons detailed below, the user will not be able to log into any websites. Lyastani et al. and others have identified analogous problems for USB security keys, particularly the difficulty of account recovery and revocation if the key is lost or stolen [1, 10, 26]. Smartphones, just like security keys, can be stolen or lost, temporarily or permanently. Identifying appropriate methods for recovering from authenticator loss is still an open problem, although FIDO2 recommends registering multiple authenticators to avoid being completely locked out [20]. As participants mentioned, though, smartphones raise additional availability issues. Unlike security keys, phones can run out of battery, making it impossible for the owner to authenticate without charging the phone. Phone availability can also be impacted by limited wireless reception. Finally, phones are also higher-value targets for theft.

Shortcomings in accessibility can also present availability challenges. For example, schemes that require biometrics to verify user identity (e.g., as might be required after confirming a push notification as we did for Neo) could cause problems for people who cannot touch a security key's capacitor, who cannot use a fingerprint scanner, or for whom facial recognition is not reliable. It is important for system designers to consider a variety of ways for users to verify their identity, potentially including some that could cause their systems to lose some of their security benefits (e.g., PINs).

When asked at the end of the study how frequently they were generally unable to access their mobile device when they needed it, a third of participants said "once a day," a third said "almost never," and the other third was "once a week" or less frequently. The variety of these results make it difficult to provide general recommendations regarding these challenges. Currently, the best approach to availability challenges may be nudging or requiring users to register multiple authenticators.

Another potential way to allow users to enjoy the security benefits of authentication methods like Neo while also considering account recovery is to enable email-based account recovery, as is typical for passwords. If a user breaks or no longer has their registered authenticator, they could receive a link in their email account to register another type of authenticator. Of course this means users have to remember at least one password, similar to using a password manager. Of course they would need to *not* be using Neo on their email account, and they would need to have a strong password for their email account. However, if they have backup unregistered authenticators at hand (e.g., old phones or platform authenticators on desktop devices), this approach could provide the usability benefits of password managers while providing the security benefits of using smartphones as roaming authenticators within FIDO2 passwordless authentication.

6 Related Work

Oogami et al. [29] conducted the first study evaluating the usability of smartphones as WebAuthn-enabled *platform authenticators*. In 2018, their website (yahoo.co.jp) was the first commercial portal to let users choose to log in from their smartphones using WebAuthn with a fingerprint. Conversely, we evaluated the use of smartphones as *roaming authenticators*. Out of their 10 participants, only three were able to complete the registration process without assistance. Although their registration process was significantly different than ours, participants in our study similarly struggled with setup.

Lyastani et al.'s [26] between-subjects lab study (N = 94) evaluated the usability of security keys with FIDO2 passwordless authentication. The authors sought to understand users' perceptions, acceptance, and concerns when using security keys for FIDO2 passwordless authentication. They found that passwordless authentication with security keys was seen as both more usable and more acceptable than passwords. Our study builds on this work, but focuses on using smartphones (instead of security keys) as roaming authenticators. While participants in their study preferred passwordless authentication with a security key to passwords, they were also concerned about account recovery and account revocation. Our participants raised these same concerns. Farke et al. [18] conducted a similar experiment in the context of a small company. Like us, they found that participants were concerned about the availability of their authenticators (security keys) in terms of physical location (e.g., losing the authenticator) and functionality (e.g., a malfunctioning authenticator).

Owens et al. [30] presented a framework for evaluating authentication schemes that use smartphones as FIDO2 roaming authenticators. They specifically highlighted user perceptions of phone availability and account recovery challenges as potential focus areas for researchers. The data we collected included the types suggested by their framework. Bonneau et al. [7] proposed a framework for evaluating web authentication schemes. This framework used 25 properties to rate 35 password-replacement schemes on usability, deployability, and security. Their expert evaluation found that no scheme analyzed offered the same benefits as passwords. Prior work has evaluated WebAuthn using this framework [13, 18, 26, 27].

To address the challenge of account recovery, Conners and Zappala [13] proposed the Let's Authenticate alternative to FIDO2 based on certificates instead of keys. Certificates are issued after users prove ownership of an account with a username and password, facilitating re-issuance if an authenticator is lost. Credential recovery and revocation problems are critical for roaming authenticators like security keys and smartphones. While Let's Authenticate eliminates the burden

of registering an authenticator with every web service, it also introduces a new trusted third party.

Klieme et al. [23] proposed an extension to FIDO2/WebAuthn that would allow continuous authentication over BLE. They created a proof-of-concept Android application to serve as a roaming authenticator and implemented a custom relying party that supported their extension. Due to browsers' lack of support for custom FIDO2/WebAuthn extensions, they *simulated* (rather than tested) extension processing by adding functionality to their custom relying party. We work around this current browser support challenge by using a Chrome browser extension to simulate Network Transport functionality.

A number of researchers have studied the usability of mobile phones as a *second* factor for authentication, including via SMS codes, TOTP codes, and push notifications [12, 15, 24, 38, 45]. Weidman and Grossklags [45] studied a transition from token-based 2FA to a push-notification 2FA system, finding that employees preferred the token-based system to the Duo app. Colnago et al. [12] studied the deployment of 2FA via the Duo app at their university. They found that 2FA adopters found it annoying, yet easy to use. Neo uses a similar push notification mechanism for authentication.

7 Limitations and Future Work

As in many user studies, our findings are somewhat limited in their generalizability by the small sample size. Additionally, in both experimental conditions, participants logged into a fictitious banking website. Consequently, they did not experience any real risk or incentive during authentication. This could cause *Password* participants to create weaker passwords than they otherwise might, and generally cause participants to behave differently than they might in real-life scenarios. We attempted to simulate risk by having participants perform simulated transactions within the banking application. However, there was no reward associated with protecting the assets in the accounts. To better simulate risk, future work could adapt the approach from Redmiles et al. [35] and assign a probability of a participant's account being "compromised" based on the characteristics of the password they created. Moreover, it is possible that users may have exhibited different behaviors or perceived things differently if the study website had a different focus (e.g., social media). Future work could explore those differences by conducting a between-subjects experiment with additional conditions that mimic other well-known web services.

Our participant pool also likely impacted our results. Because users with prior 2FA experience with push notifications are over-represented (45% vs. 19% in the general US population, according to Engler [17]), and we found that this prior experience made participants view Neo as more usable, the results from this study could be seen as overly optimistic. Although MTurk users are often more diverse in terms of age, income, education level, and geography than traditional social science pools, they are also younger, Whiter, and more tech-savvy than the general US population [34]. Because we required that participants live in the US (to reduce confounding factors), our results are not reflective of global populations. Future work should study more diverse populations.

As previously discussed, *Neo* overall had a far greater attrition rate than *Password* despite random assignment. We listed several potential causes in Section 3.3. Some amount of the dropout was likely a result of the difficulty associated with setting up Neo. Thus, our final set of participants may be biased and present overly optimistic results. However, if this effect were strongly present, one might expect Neo to have been found to be more usable than passwords. We found the opposite. Future work should study the setup process and daily use separately, even more closely tracking attrition. We also observed a jump in the *Neo* authentication success rate from Session 1 to 2. We speculate that this jump reflects the learning effect for a new system, although we cannot speak definitively about what aspects were barriers in Session 1. Studying this increase is an avenue for future work.

We tested only a simple password-composition policy. Future work should test a variety of different policies and separately test password managers to further understand how WebAuthn compares to various password use cases. While our participants were Chrome users, our study platform did not enforce the use of Google Chrome when registering or during authentication. This means that participants in *Password* could have used other browsers, introducing a confound. Finally, future work should study using platforms like Neo across multiple websites. A user has to register separately each time they want to add Neo as an authenticator on a new website; we only studied its usability on a single website.

8 Conclusion

We conducted a between-subjects (N = 97), longitudinal study of FIDO2 passwordless authentication with smartphones as roaming authenticators. Participants recognized the security benefits of the Neo smartphone-based passwordless authentication scheme, yet still found passwords to be more usable. Nonetheless, many participants were willing to use Neo over passwords for five of the six account types we asked about. Participant were acutely aware of challenges associated with losing an authenticator and stressed the need for account recovery methods. Participants suggested that the setup process for Neo be simplified, that different ways of verifying user presence (e.g., PIN, facial recognition) be made available for authentication, and that account recovery/backup methods be added. While some of the concerns participants had (e.g., setup issues) were unique to the design of Neo, we believe our findings highlight issues and opportunities designers of smartphone-based FIDO2 passwordless authentication must consider when implementing new schemes.

Acknowledgments

We would like to thank our anonymous shepherd and reviewers for their helpful feedback. We would also like to thank multiple people at Duo Security for their support of this project, including Bronwyn Woods, Jeremy Erickson, Nick Mooney, Nick Steele, Rich Smith, Brian Lindauer, and the entire Data Science Team.

References

[1] Seb Aebischer, Claudio Dettoni, Graeme Jenkinson, Kat Krol, David Llewellyn-Jones, Toshiyuki Masui, and Frank Stajano. Pico in the Wild: Replacing Passwords, One Site at a Time. In *Proc. EuroUSEC*, 2017.

[2] Eldridge Lee Alexander, James Leslie Barclay, Nicholas James Mooney, and Mujtaba Hussain. Identity Services for Passwordless Authentication. US Patent US20200403993A1, December 2020.

[3] FIDO Alliance. FIDO2: WebAuthn & CTAP, May 2020. https://web.archive.org/web/20200512070231/https://fidoalliance.org/fido2/.

[4] Aaron Bangor, Philip Kortum, and James Miller. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *Journal of Usability Studies*, 4(3):114–123, 2009.

[5] Lujo Bauer, Cristian Bravo-Lillo, Elli Fragkaki, and William Melicher. A Comparison of Users' Perceptions of and Willingness to Use Google, Facebook, and Google+ Single-sign-on Functionality. In *Proc. DIM*, 2013.

[6] Chandrasekhar Bhagavatula, Blase Ur, Kevin Iacovino, Su Mon Kywe, Lorrie Faith Cranor, and Marios Savvides. Biometric Authentication on iPhone and Android: Usability, Perceptions, and Influences on Adoption. In *Proc. USEC*, 2015.

[7] Joseph Bonneau, Cormac Herley, Paul C. van Oorschot, and Frank Stajano. The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes. In *Proc. IEEE S&P*, 2012.

[8] Dhiman Chakraborty and Sven Bugiel. simFIDO: FIDO2 User Authentication with simTPM. In *Proc. CCS*, 2019.

[9] Dhiman Chakraborty, Lucjan Hanzlik, and Sven Bugiel. simTPM: User-centric TPM for Mobile Devices. In *Proc. USENIX Security*, 2019.

[10] Stéphane Ciolino, Simon Parkin, and Paul Dunphy. Of Two Minds about Two-Factor: Understanding Everyday FIDO U2F Usability through Device Comparison and Experience Sampling. In *Proc. SOUPS*, 2019.

[11] Jacob Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

[12] Jessica Colnago, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Lorrie Cranor, and Nicolas Christin. "It's not actually that horrible": Exploring Adoption of Two-Factor Authentication at a University. In *Proc. CHI*, 2018.

[13] James S. Conners and Daniel Zappala. Let's Authenticate: Automated Cryptographic Authentication for the Web with Simple Account Recovery. In *Proc. WAY*, 2019.

[14] Sanchari Das, Andrew Dingman, and L. Jean Camp. Why Johnny Doesn't Use Two Factor: A Two-Phase Usability Study of the FIDO U2F Security Key. In *Proc. FC*, 2018.

[15] Emiliano De Cristofaro, Honglu Du, Julien Freudiger, and Greg Norcie. A Comparative Usability Study of Two-Factor Authentication. In *Proc. USEC*, 2014.

[16] Department of Health and Human Services. System Usability Scale (SUS), Sep 2013. https://web.archive.org/web/20200201012855/https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html.

[17] Maggie Engler. 2019 State of the Auth: Experiences and Perceptions of Multi-Factor Authentication. Duo Security E-book, December 2019. https://duo.com/assets/ebooks/state-of-the-auth-2019.pdf.

[18] Florian M. Farke, Lennart Lorenz, Theodor Schnitzler, Philipp Markert, and Markus Dürmuth. "You still use the password after all"–Exploring FIDO2 Security Keys in a Small Company. In *Proc. SOUPS*, 2020.

[19] Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, 2013.

[20] Hidehito Gomi, Bill Leddy, and Dean H. Saxe. Recommended Account Recovery Practices for FIDO Relying Parties. *FIDO Alliance*, 2019. https://web.archive.org/web/20210520070746/https://fidoalliance.org/recommended-account-recovery-practices/.

[21] Troy Hunt. Passwords Evolved: Authentication Guidance for the Modern Era, 2020. https://web.archive.org/web/20200501185526/https://www.troyhunt.com/passwords-evolved-authentication-guidance-for-the-modern-era/.

[22] Ryan Kennedy, Scott Clifford, Tyler Burleigh, Ryan Jewell, and Philip Waggoner. The Shape of and Solutions to the MTurk Quality Crisis. *SSRN*, 2018. https://www.ssrn.com/abstract=3272468.

[23] Eric Klieme, Jonathan Wilke, Niklas van Dornick, and Christoph Meinel. FIDOnuous: A FIDO2/WebAuthn Extension to Support Continuous Web Authentication. In *Proc. TrustCom*, 2020.

[24] Kat Krol, Eleni Philippou, Emiliano De Cristofaro, and M. Angela Sasse. "They brought in the horrible key ring thing!" Analysing the Usability of Two-Factor Authentication in UK Online Banking. In *Proc. USEC*, 2015.

[25] Leona Lassak, Annika Hildebrandt, Maximilian Golla, and Blase Ur. "It's Stored, Hopefully, on an Encrypted Server": Mitigating Users' Misconceptions About FIDO2 Biometric WebAuthn. In *Proc. USENIX Security*, 2021.

[26] Sanam Ghorbani Lyastani, Michael Schilling, Michaela Neumayr, Michael Backes, and Sven Bugiel. Is FIDO2 the Kingslayer of User Authentication? A Comparative Usability Study of FIDO2 Passwordless Authentication. In *Proc. IEEE S&P*, 2020.

[27] Robbie MacGregor. Evaluating the Android Security Key Scheme: An Early Usability, Deployability, Security Evaluation with Comparative Analysis. In *Proc. WAY*, 2019.

[28] Nick Mooney. Addition of a Network Transport, 2020. https://github.com/w3c/webauthn/issues/1381.

[29] Wataru Oogami, Hidehito Gomi, Shuji Yamaguchi, Shota Yamanaka, and Tatsuru Higurashi. Poster: Observation Study on Usability Challenges for Fingerprint Authentication Using WebAuthn-enabled Android Smartphones. In *Proc. SOUPS Posters*, 2020.

[30] Kentrell Owens, Blase Ur, and Olabode Anise. A Framework for Evaluating the Usability and Security of Smartphones as FIDO2 Roaming Authenticators. In *Proc. WAY*, 2020.

[31] Sarah Pearman, Shikun Aerin Zhang, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Why People (Don't) Use Password Managers Effectively. In *Proc. SOUPS*, 2019.

[32] Pew Research Center. Demographics of Mobile Device Ownership and Adoption in the United States, 2019. https://web.archive.org/web/20210606233708/https://www.pewresearch.org/internet/fact-sheet/mobile/.

[33] Suby Raman. Guide to Web Authentication, 2021. https://webauthn.guide.

[34] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples. In *Proc. IEEE S&P*, 2019.

[35] Elissa M. Redmiles, Michelle L. Mazurek, and John P. Dickerson. Dancing Pigs or Externalities? Measuring the Rationality of Security Decisions. In *Proc. EC*, 2018.

[36] Elissa M. Redmiles, Noel Warford, Amritha Jayanti, Aravind Koneru, Sean Kross, Miraida Morales, Rock Stevens, and Michelle L. Mazurek. A Comprehensive Quality Evaluation of Security and Privacy Advice on the Web. In *Proc. USENIX Security*, 2020.

[37] Ken Reese. 2FA Banking Website. https://bitbucket.org/isrlauth/ken-bank-thesis/src/master/, 2020.

[38] Ken Reese, Trevor Smith, Jonathan Dutson, Jonathan Armknecht, Jacob Cameron, and Kent Seamons. A Usability Study of Five Two-factor Authentication Methods. In *Proc. SOUPS*, 2019.

[39] Joshua Reynolds, Trevor Smith, Ken Reese, Luke Dickinson, Scott Ruoti, and Kent Seamons. A Tale of Two Studies: The Best and Worst of YubiKey Usability. In *Proc. IEEE S&P*, 2018.

[40] Scott Ruoti, Brent Roberts, and Kent Seamons. Authentication Melee: A Usability Analysis of Seven Web Authentication Systems. In *Proc. WWW*, 2015.

[41] Aaron Smith. Americans and Cybersecurity. Pew Research Center, January 2017. https://web.archive.org/web/20210514203628/https://www.pewresearch.org/internet/2017/01/26/2-password-management-and-mobile-security/.

[42] San-Tsai Sun, Eric Pospisil, Ildar Muslukhov, Nuray Dindar, Kirstie Hawkey, and Konstantin Beznosov. What Makes Users Refuse Web Single Sign-on? An Empirical Investigation of OpenID. In *Proc. SOUPS*, 2011.

[43] Joshua Tan, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Practical Recommendations for Stronger, More Usable Passwords Combining Minimum-strength, Minimum-length, and Blocklist Requirements. In *Proc. CCS*, 2020.

[44] W3C. Web Authentication, 2019. https://www.w3.org/TR/webauthn/.

[45] Jake Weidman and Jens Grossklags. I Like It, But I Hate It: Employee Perceptions Towards an Institutional Transition to BYOD Second-Factor Authentication. In *Proc. ACSAC*, 2017.

[46] Yubico. User Presence vs. User Verification, 2021. https://web.archive.org/web/20210605113506/https://developers.yubico.com/WebAuthn/WebAuthn_Developer_Guide/User_Presence_vs_User_Verification.html.

[47] Yubico. Works with YubiKey Catalog (FIDO2), 2021. https://www.yubico.com/works-with-yubikey/catalog/#protocol=fido2&usecase=all&key=all.

[48] Tin Zaw and Richard Yew. 2017 Verizon Data Breach Investigations Report (DBIR) from the Perspective of Exterior Security Perimeter. Verizon Media Platform, 2017. https://web.archive.org/web/20200409012027/https://www.verizondigitalmedia.com/blog/2017-verizon-data-breach-investigations-report/.

A Screening Survey

[This survey was sent to MTurkers who accepted our HIT.]

Please complete the following 1 minute screening survey to see if you qualify for a two-week longitudinal study on online authentication. This study may require you to install a mobile application and/or a Chrome extension. You will be asked to login into a web application ten times over the course of two weeks, completing a simple task each time. Completing all of the tasks over two weeks should not take more than 75 minute total. After completing the first task you will take an initial survey. After two weeks pass, we will send you a link to a final survey. Upon adequate completion of the final survey, you will receive your $30 in compensation as a bonus.

Do not take this survey unless you meet the following criteria:

- Have an Android phone with a fingerprint sensor
- Have Android version 9.0+ (to check what version of Android you have, go to your phone's "Settings," and search "Android version," or you can visit the following website from your mobile phone: https://whatismyandroidversion.com/)
- Have Google Chrome installed on your computer
- Are an adult currently living in the USA

If you take this screening survey and do not meet the above criteria, you will not receive compensation.

If you qualify for the survey, we will message you with more details about the study and send the $30 as a bonus upon completion of the longitudinal study.

By taking the survey, you are agreeing to the non-disclosure agreement found at the following link: https://bankoferie.com/nda ◯ I agree ◯ I do not accept and will not participate in this study

Please enter your MTurk ID below. Please ensure that it is correct to ensure that we are able to compensate you for your participation. _____

Do you have an Android mobile device? ◯ Yes ◯ No

If you answered yes to the above question, what Android software version do you have do have? To check this, go to your phone's "Settings," and search "Android version" or you can visit https://whatismyandroidversion.com/ from your mobile phone. _____

Are you an adult (18+ years old) currently living in the United States? ◯ Yes ◯ No

Do you have Google Chrome installed on your computer? ◯ Yes ◯ No

Does your Android phone have a fingerprint sensor? ◯ Yes ◯ No

B Survey Instrument

[Below we highlight the difference between the initial/exit surveys received by participants in the Neo and passwords conditions after the completion of their tasks.]
Please enter your MTurk ID below. Please ensure that it is correct to ensure that we are able to compensate you for your participation. _____

Please enter the username that you used for the study. Please ensure that it is correct to ensure that we are able to compensate you for your participation. _____

B.1 System Usability Scale

In the following survey, the word "system" refers to the [mobile phone-based or passwords-based] authentication method you used to log into your account. Please state your level of agreement or disagreement for the following statements based on your experience with this system. There are no right or wrong answers.

[Response choices: ◯ Strongly agree ◯ Agree ◯ Neither agree nor disagree ◯ Disagree ◯ Strongly disagree]
Questions:

- I think that I would like to use this system frequently.
- I found the system unnecessarily complex.
- I thought the system was easy to use.
- I think that I would need the support of a technical person to be able to use this system.
- I found the various functions in this system were well integrated.
- I thought there was too much inconsistency in this system.
- I would imagine that most people would learn to use this system very quickly.
- I found the system very awkward to use.
- I felt very confident using the system.
- I needed to learn a lot of things before I could get going with this system.

B.2 Additional questions

These questions are about your experience with [setup or day-to-day use] of the [mobile phone-based or passwords-based] method you used to log into your account in the web application. Please answer them thoroughly and honestly. There are no right or wrong answers.

How would you describe your general experience with the authentication method you used? _____

What advantages do you see with using this authentication method? _____

What disadvantages do you see with using this authentication method? _____

[Final, Passwords only] Do you think using this authentication method is the best available for protecting the safety of your online accounts? _____

[Final, Neo only] If you were to recommend Neo to a friend, how would you describe its benefits? _____

[Neo only] How likely are you to choose Neo over passwords for the following types of accounts, if Neo were widely available?

[Final, Neo only] What changes would need to be made to Neo to make you more likely to use it? _____

[Final, Neo only] Did you previously visit the website (https://webauthn.guide) mentioned in the tutorial to learn more about WebAuthn? ◯ Yes ◯ No ◯ I don't remember

	Very likely		Neutral		Not likely	N/A.
Dating services (e.g. Bumble, OkCupid)	◯	◯	◯	◯	◯	◯
Streaming services (e.g. Netflix, Hulu)	◯	◯	◯	◯	◯	◯
Social media	◯	◯	◯	◯	◯	◯
Healthcare services	◯	◯	◯	◯	◯	◯
Bank	◯	◯	◯	◯	◯	◯
Email	◯	◯	◯	◯	◯	◯

[Final, Neo only] What other sources, if any, did you use to learn about WebAuthn (if you didn't use any input N/A)? _____

[Final, Neo only] Do you think using this authentication method makes an online account safer? _____

Generally, how frequently have you not been able to access your mobile phone when you needed it? ◯ Once per day ◯ Once per week ◯ Once per month ◯ Once per year ◯ Almost never ◯ Other _____

[Initial only] How do you typically choose your password for a new email account? ◯ Reuse an existing password ◯ Modifying an existing password ◯ Create an entirely new password on my own ◯ Randomly generate an entirely new password with browser/password manager/other tool ◯ I prefer not to answer ◯ Other _____

[Final, Neo only] Did you have fingerprint enabled on your Android phone PRIOR to beginning this study? ◯ Yes◯ No ◯ Other _____

[Final only] Anything else you'd like to add? _____

[Initial only] Have you ever been a victim of account compromise/hacking? ◯ Yes (please briefly describe the incident)_____ ◯ No

B.3 Demographic Info

[Initial only] Please choose the range that includes your age. ◯ 18-24 years old ◯ 25-34 years old ◯ 35-44 years old ◯ 45-54 years old ◯ 55-64 years old ◯ 65-74 years old ◯ 75+ years old

[Initial only] Please choose your race/ethnicity (select all that apply). ◯ American Indian or Alaska Native ◯ Asian ◯ Black or African American ◯ Hispanic ◯ Native Hawaiian or Other Pacific Islander ◯ White ◯ Other _____

[Initial only] What is your gender? _____

[Initial only] Please indicate if you have a computer science background. ◯ Yes ◯ No

[Initial only] Please indicate your highest educational degree. ◯ High School Diploma/GED ◯ Some college but no degree ◯ Associate's degree ◯ Bachelor's degree ◯ Professional degree (e.g. Master's, PhD, MD, JD) ◯ Other _____

[Initial only] What forms of two-factor authentication have you used in the past, if any? ◯ SMS/Text Message ◯ TOTP code generator app (e.g. Google Authenticator, Authy, DUO Mobile) ◯ Pre-generated codes (that you printed or wrote down to use later) ◯ Push notification based mobile app (e.g. Google Prompt, Authy OneTouch, DuoMobile) ◯ Physical security keys (e.g. YubiKey, Titan) ◯ Other _____

C Selected screenshots from Neo setup guide

We made a setup guide to help participants successfully register and authenticate using Neo. Participants were required to pair the mobile application with their browser to share a secret via a QR code.

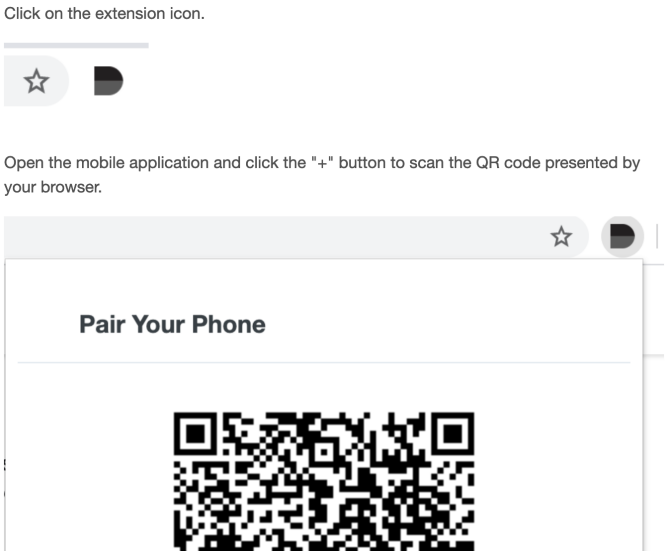


Figure 5: A screenshot of the Chrome extension during account registration from the Neo setup guide.

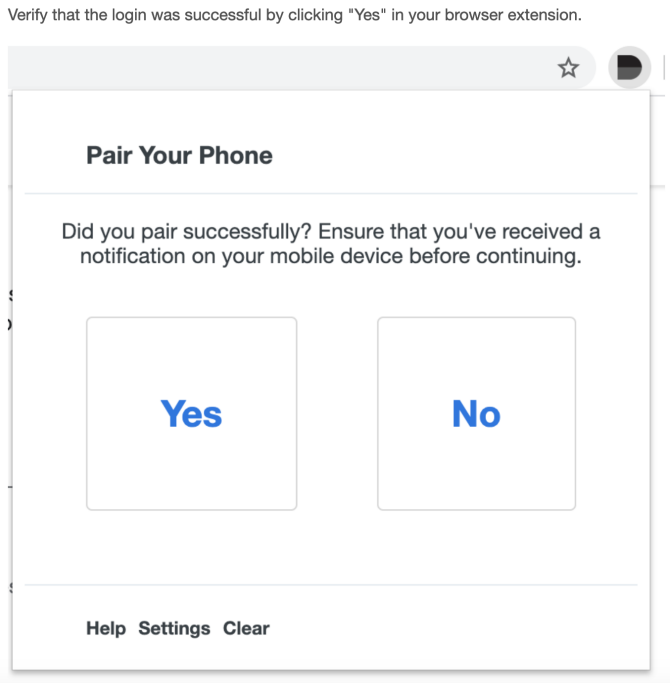


Figure 6: A screenshot of the Chrome extension prompting a user to confirm that the pairing was successful.

D Supplemental Material

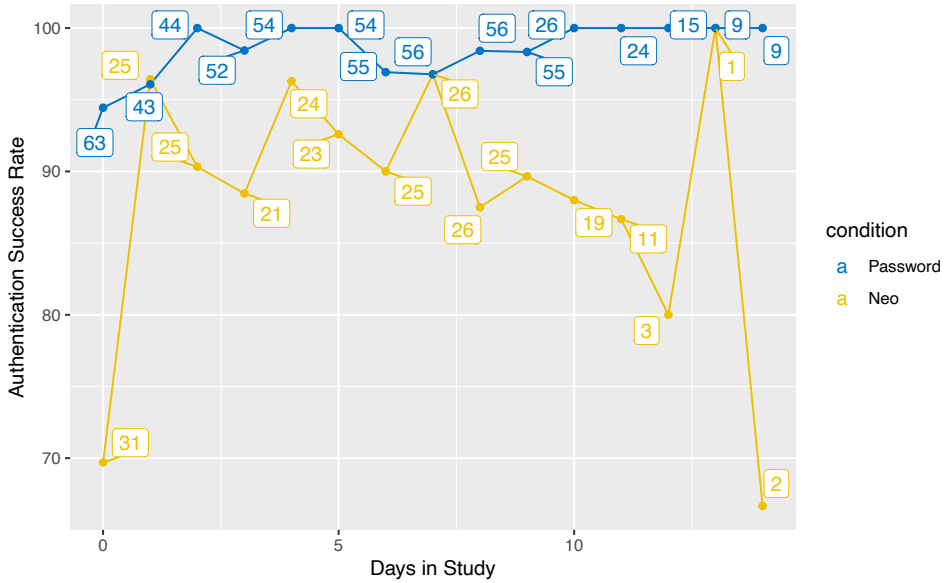


Figure 7: Aggregate authentication success rate over time by condition. The labels indicate the number of unique participants who authenticated on that day in the specified condition. We required that participants log in on ten days within a fourteen day window, and many participants simply logged in for the first ten days of the study. Note that the number of authentication *attempts* (reflected in the authentication success rate) can be greater than the number of participants in the case of failed authentication attempts. For instance, on Day 14 there were two Neo participants, one of whom logged in successfully on the first attempt and one of whom had a failed authentication attempt followed by a successful attempt.

Never ever or no matter what: Investigating Adoption Intentions and Misconceptions about the Corona-Warn-App in Germany

Maximilian Häring*
University of Bonn

Eva Gerlitz*
Fraunhofer FKIE

Christian Tiefenau*
University of Bonn

Matthew Smith
University of Bonn, Fraunhofer FKIE

Dominik Wermke
CISPA, University of Hannover

Sascha Fahl
CISPA, University of Hannover

Yasemin Acar
Max Planck Institute for Security and Privacy

Abstract

To help tackle the COVID-19 pandemic, the tech community has put forward proximity detection apps to help warn people who might have been exposed to the coronavirus. The privacy implications of such apps have been discussed both in academic circles and the general population. The discussion in Germany focused on the trade-off between a centralized or decentralized approach for data collection and processing and their implications. Specifically, privacy dominated the public debate about the proposed "Corona-Warn-App." This paper presents a study with a quota sample of the German population ($n = 744$) to assess what the population knew about the soon-to-be-released app and their willingness to use it. We also presented participants potential properties the app could have and asked them how these would affect their usage intention. Based on our findings, we discuss our participants' views on privacy and functionality, including their perception of selected centralized and decentralized features. We also examine a wide range of false beliefs and information that was not communicated successfully. Especially technical details, such as that the app would use Bluetooth, as opposed to location services, were unknown to many participants. Our results give insights on the complicated relationship of trust in the government and public communication on the population's willingness to adopt the app.

1 Introduction

Since the spread of COVID-19 in 2020, governments have been developing measures to fight its transmission. One of these measures is the use of contact tracing apps. Early in 2020, the public media in Germany discussed two different approaches. The *centralized app* was based on PEPP-PT [24], while the *decentralized app*, was based on DP-3T [44]. Both approaches come with advantages and disadvantages. The public debate was driven by researchers who signed an open letter (April 2020) backing the decentralized approach [7], as well as privacy advocates (April 2020) [11]. A major argument was that the general population would only be willing to adopt the app in sufficient numbers if privacy was preserved [7]. The German government had previously committed to the centralized app, which they abandoned during development due to the public debate, starting a new development project based on the decentralized approach at the end of April. The media extensively discussed this decision, and the government, via direct appeals and public media, encouraged people to install the app. In this context, we were interested in finding out how much the general public understood about the newly announced but, at the time of conducting the study, yet to be released decentralized app. We were also interested in the general public's attitudes towards potential properties, particularly those about the centralized and decentralized approaches' advantages and disadvantages. To gain insights into these issues, we conducted an online survey study from May 30 to June 11, 2020, with a quota sample of 744 participants from Germany. The app was released on June 16 and became one of the most installed European apps [43].

In this paper, we make the following contributions:

- We conducted the first study to assess participants' knowledge of and beliefs about the planned Corona-Warn-App (CWA) after the app features were published and broadly discussed in the media. This is in contrast to other studies in Germany that focused on hypothetical apps.
- We assess *how accurately* participants could identify the properties of the planned German contact tracing app.
- The German public discourse was dominated by the discussion of a centralized versus decentralized application.

* These main authors contributed equally to this work.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

We offer insights into the level of relevance of the app's capabilities linked to the centralized approach.

- We compare our work to contemporary work that assessed willingness to install various hypothetical tracing apps in Germany [68].

2 Related Work

In this section, we discuss related work in three areas relevant to the context of this work. 1) The discussion about technical aspects of tracing apps and their history/development in Germany, in particular between the centralized and decentralized approach; 2) studies about the acceptance of contract tracing apps and influence of factors in released or hypothetical corona tracing apps. 3) Research on the existing knowledge of users about upcoming or already released corona tracing apps.

We note that most of the literature was published after we designed and conducted our study, so we compare our results retrospectively.

2.1 Technical background/history

The idea of supporting contact tracing with mobile apps emerged early in the pandemic. The first working app to fight COVID-19 was released in March 2020 in Singapore [21], only two months after the first reported infections outside of China [35]. The app had the disadvantage that it had to be constantly the visible app on the smartphone to allow data exchange. For such an app to effectively support health agencies in contact tracing, a large set of the population has to use it, depending on the overall scenario (e.g., how quarantine is handled) [55].

In Germany, two approaches to collect and process the data were discussed. In the centralized version called "Pan-European Privacy-Preserving Proximity Tracing" (PEPP-PT) [24] all collected encounters, namely contact-ID and timestamp of encounters with other app users would be uploaded and stored on a central server. In the decentralized version "Decentralised Privacy-Preserving Proximity Tracing" (DP-3T) [44], all encounters remain on the users' smartphone. If a user tests positive for COVID-19, they can upload all their cryptographic keys (from which the IDs can be derived) to a server. Once a day, a list with keys of people who reported their positive COVID-19 tests is downloaded to all users' smartphones and compared to locally stored encounters within the last 14 days. The important difference is that determining whether a user has been at risk of contracting COVID-19 is calculated on their smartphone itself and the encounter data never leaves the phone. At the beginning of the discussion, the German government wanted to follow the centralized approach [25]. After two open letters in April 2020 suggesting the usage of DP-3T [7, 23], the German government changed course and pivoted to the decentralized approach on April 26 [12]. Two days later, a press release was published that contained (technical) information about the app, such as that it would work with Bluetooth [27].

2.2 User acceptance of tracing-apps

Since the idea to use apps that would support the contact tracing work of the health departments to contain COVID-19 became popular among governments worldwide, researchers aimed at understanding user preferences to allow for broad adoption. Studies were conducted in Australia [66], Europe [65] (including Belgium [70], France [45, 52], Germany [45, 48, 56, 59, 60, 67, 68], Italy [45], Ireland [63], Switzerland [69] and the UK [45, 46, 57, 71, 72]) and the USA [41, 45, 51, 53, 54, 58, 59, 61, 62, 64, 65, 68, 73]. As Utz et al. [68] and Kostka et al. [59] found similarities for Europe and America, we focus on work conducted with those populations.

Most conducted studies were choice-based conjoint experiments, in which participants were asked to select which app of several they would prefer or were given different app configurations for which they had to decide if they would install such an app [48, 68, 71, 73]. Some studies asked to imagine a corona tracing app has already been released [45, 54, 57, 59, 61, 70] We are aware of only a few studies that looked at the user acceptance and influencing factors on the acceptance for the app that was already launched in the surveyed country [66, 69].

Investigated Factors The studies explored factors that could influence participants' intention to install and use the corona tracing app. Several authors investigated how personal characteristics, such as demographics or one's experience with the pandemic, impacted the acceptance of corona tracing apps [45, 46, 48, 52–54, 56, 59, 61, 68, 69, 73]. Amongst others, people who were male [56, 61, 68], had higher trust in the respondent's government [45, 48, 52, 68, 69], health authorities [69] and others in general [56], had higher income [61, 69] or lived in urban areas [59, 61] were more likely to install a tracing app. While some authors noted that younger participants were more inclined to use a tracing app [53, 56, 61], others found the opposite [54, 59]. Looking at pandemic-related factors, health concerns during the pandemic, and personal experience with COVID-19 increased the willingness to install a tracing app [48, 52, 53, 59, 68]. Additionally, better adherence to COVID-19 regulations was a positive influence [69]. Fear and anxiety concerning changes in government rules [46] impacted participants negatively.

Apart from factors that might influence the acceptance of contact tracing apps in general, many studies were conducted to find which app design choices would be considered positively or negatively by participants. The studies covered different attributes (e.g. what data will be collected) [48, 58, 61, 62, 73], the apps purpose [68] or what institution will develop, host, distribute or own the app [41, 48, 53, 57, 64, 65, 71]. Li et al. [62] found a preference

for the centralized, while Zhang et al. [73] had more participants who were willing to use a decentralized app. Horvath et al. [57] found a centralized national health system to be favored. Participants rated health agencies more trustworthy than their government as a whole concerning corona tracing apps [57, 65]. Still, Simko et al. [65] found no entity that everyone trusted. Anonymous data collection impacted participants positively in their decision to use an app [48] and it was perceived negatively if the collected data can uniquely identify individuals [68]. Independent of app design choices, studies often found a subset of participants who did not like any of the proposed apps [58, 62, 68].

Aside from app properties, researchers looked into effects an app could have, and the influence this has on adoption [48, 61, 71] such as malfunctions of the app [58, 68] or the perceived effectiveness in fighting against COVID-19 [59–61, 70]. They found that participants' perception of the (public) health benefits an app would offer and other people's willingness to use it explained the usage intention better than app design choices and personal characteristics [61]. Performance expectancy and the benefit were also among the most critical predictors in other studies [59, 60, 70]. Malfunction in contact tracing was found to be of negative influence [68] and participants valued false negatives worse than false positives [58]. The willingness to use contact tracing apps increased if its usage is linked to priority testing [48, 71]

Further, numerous studies identified the primary reason why users would or would not install tracing apps [45, 46, 56, 63, 66, 68, 69, 72]. In their studies, privacy concerns [45, 46, 56, 66, 68, 69, 72], technical concerns or lack of technical equipment [56, 66, 69], distrust in the government [66] or the fear of surveillance at the end of the pandemic [45, 63] and doubts about the effectiveness or benefit [56, 69] were brought up as negative influences. The following topics were mentioned as reasons for using a tracing app: willingness to protect family and friends [45, 63], a sense of responsibility for the community [45, 63, 72] and the hope that the app may stop the pandemic [45].

2.3 Knowledge about corona tracing apps

The subsequent studies examined what participants knew about corona tracing apps apart from factors and properties influencing users' installation or usage intention. Simko et al. [65] conducted surveys for seven months in the US and Europe, focusing on contact tracing and privacy and asking for potential app properties. Within the participants' answers, they identified several false mental models, e.g., that proximity tracing is less secure than location tracking due to constant communication between devices. Zhang et al. [73] surveyed 2000 participants in the USA to measure the support for nine different COVID-19 surveillance measures, including tracing apps. While analyzing, they noticed participants had many misunderstandings about the described app, although the description was still visible when they answered the questions. For example, a third believed they would receive the names of infected people they had been in contact with. The number of incorrect answers could not predict the participants' usage intention. Williams et al. [72] conducted focus groups in the UK to explore public attitudes to the proposed contact tracing app. The authors found the most common misconception was that the app would make it possible to precisely identify COVID-19 cases in their vicinity and amongst their contacts. In one study that took place outside Europe and America, Thomas et al. [66] surveyed 1500 Australians after the national tracing app was released and examined participants' knowledge about it. Around 70% knew the app would make it easier and faster to inform people exposed to COVID-19 and warn users who would not have been warned otherwise. However, 50% did not reject the assumption that their personal information would be used after the pandemic, and 57.4% believed the app would warn if infected people were near them.

3 Methodology

This section describes instrument development and the conducted survey, our recruitment, and the data analysis process.

3.1 Survey Development

We followed the public discussion of the CWA. We were interested in the information that potential users have, mainly as discussions focused on whether enough people would install it and why (not). Much of this discussion in Germany revolved around the topic "centralized versus decentralized" and the claim that this would heavily influence the willingness to install. As there was little concrete related work on users' perception, we were also interested in the broader topic of acceptance and beliefs. We discussed factors and potential influences with other researchers and iterated multiple times over the survey.

Pre-Testing Before handing out the survey, we conducted several test rounds with colleagues who were not involved in the survey creation to identify comprehension problems. Following that, we asked 19 computer science students to fill the survey and provide additional feedback about unclear sections and inconsistencies. After this, we additionally sampled 50 participants on Clickworker [6]. Finally, we asked five participants without a technical background to fill the survey while thinking aloud. Before starting the final study, Qualtrics [29] additionally sampled 50 people. This pilot study helped get an overview of the duration, evaluate the randomization and spot flaws in the survey logic.

3.2 Survey Content

To inform the survey structure and questions, we looked at the different available approaches to develop a contact tracing

app and followed the media discussion. The final survey consisted of the following described four parts and can be seen in Appendix A.

Media Sources and Knowledge In the first part, we asked whether the participants had already heard of the planned app and asked for their knowledge sources (e.g., public broadcasters, family members, social media, or official government websites) (Q7). After this, we asked questions that assess their knowledge of the properties of the app in general (Q8). We also asked such questions for two scenarios: what happens if other users are infected (Q9) or if the users themselves are infected (Q10). In these three question blocks, we included 23 statements that were either correct (8 statements) or incorrect concerning the soon-to-be-released app (15 statements). As incorrect statements, we used properties of another 'corona app' released in Germany [8][1] or were discussed in media at the time of the survey. For example, we included the misconception that the app will share all phone numbers saved on the user's phone or share a movement profile with the government. Three statements were neither correct nor incorrect for the released app. Details of all these statements can be found in Table 6.

Disposition to use In the second part of the survey, we showed the participants a minimal description of the app, including the information that its primary purpose will be to warn users who have been close to infected persons and use Bluetooth to detect other app-users. Following that, the participants were asked whether they are planning to use the app, using a question with five possible answers ranging from "1 - Definitely will use it" to "5 - Definitely will not use it" (Q12). We also asked to report their primary reason for their choice in a text field (Q13).

Potential Properties The third part presented 23 hypothetical statements, from now on called *potential properties*, about the app (Q14). The participants were asked how these statements would affect their willingness to use the app if the app would work this way (5 answer options, from "1 - Definitely would use it" to "5 - Definitely would not use it"). In this section, we added an attention check question. Six of those properties can be attributed to a centralized approach, while one would only be valid for the CWA app that is based on the decentralized approach. Additionally, 12 properties were correct for the to-be-released app, while 11 were incorrect. Details of all the presented statements can be seen in Table 7.

Demographics In the end, we asked for demographic data and how COVID-19 impacted their lives (Q16-29).

[1]The app can be used to share fitness data with the RKI.

3.3 Recruitment

We used Qualtrics [29] to recruit a representative German sample according to age, education, household income, and federal state/region. Qualtrics provided representative numbers for age, education, and region, numbers for income were taken from the Federal Statistical Office of Germany from 2017 [17]. Due to the nature of online surveys, older participants were underrepresented, and we could not entirely fulfill our quotas for a representative sample. The final distribution after sanitizing the data together with our targeted quotas can be seen in Table 1. The study was conducted from May 30 to June 11; thus, shortly before the app was launched on June 16, 2020. To take part in the study, it was required that the participants owned a smartphone since that is a precondition to use the app. 1025 participants took part in the study for which we paid Qualtrics €4000.

3.4 Data quality

During the study, Qualtrics excluded participants that 1) took less than half the median of the time the participants needed in the final pilot study (243 seconds) for completing the survey, [2] or 2) failed the attention check question in the potential properties question block.

To ensure our participants were paying attention, we included a straightforward attention check (Q14) and one comprehension check question (Q11). The comprehension check question gave a short explanation of how the app will work (specifically mentioning using Bluetooth for contact tracing). It then asked what technology the app will utilize for contact tracing. We excluded participants from our analysis who did not choose "Bluetooth".

When we designed this question, it seemed quite straightforward. To our surprise 262 participants failed this question. We then discussed whether we had overlooked genuine reasons why this question might be answered incorrectly. Potentially, participants who read our description text did not believe it and answered true to their previous or internal beliefs. It is also possible that our description was too complex for some to understand and thus could mean that they misunderstood other questions.

We also discussed the possibility of excluding participants due to inconsistent or odd answers, e.g., a participant stating that they are a civil servant but also stating that they lost their job[3] or stating that the app used Bluetooth in one question and stating otherwise in another. However, after an in-depth discussion, we decided against this. We looked at the free text answers of participants who had such inconsistencies, but we found them generally to be as plausible as those who did not and did not find any other warning markers.

[2]This is a standard procedure at Qualtrics; we do not know how many participants were excluded.

[3]In Germany this combination is incredibly rare

3.5 Analysis

We analyzed the data in two different ways. Most of the results concern a quantitative analysis of the answers. One free text answer was analyzed qualitatively. Percentages are reported rounded.

Quantitative For our quantitative evaluation, besides reporting, we performed an ordered logit regression with model selection, an ordered logit regression model containing all potential properties, and hypothesis testing. For the app usage intention, we decided to combine participants who answered "I don't know" and those who answered "I am undecided".

In the *Media Sources and Knowledge*-section (Q8-10) of the survey, we asked participants whether they thought the presented statements were correct for the CWA. False statements required no click from the participant to give the correct answer. This may influence the measured correctness of their beliefs, besides the point that some statements may be easier or harder to know. We, therefore, only report true statements that were known as (positive) knowledge and false statements that were clicked as false beliefs.

Coding process Participants were asked to indicate their primary reason for wanting to use or not use the CWA (Q13). One researcher looked at the answers and coded them according to the participant's misconceptions. All presented quotes were discussed and agreed upon by two researchers. All quotes were originally in German and translated into English by the authors.

Regressions For our exploratory regression model, we conduct a model selection approach by computing a set of candidate models based on different factor combinations, and selecting the final model based on a combination of the best Akaike Information Criterion (AIC) [49] and Bayesian Information Criterion (BIC) [1]. Possible factor categories and corresponding baselines are reported in Table 5.

Our final ordered logit regression (cf. Table 3) reports change as log odds to highlight trends: a negative value directly correlates to a negative effect and vice versa for positive values. In addition, we report a 95% confidence interval (C.I.) and a p-value. For convenience, we highlight factors below an arbitrary significance cut-off of 0.05 with an asterisk ($*$).

In addition, to investigate potential effects of different app features, we conducted an ordered logit regression (cf. Table 4) with all app features as factors.

3.6 Ethics

Our study was reviewed and approved by our institution's Research Ethics Board. We also adhered to the German data protection laws and the GDPR in the EU. For all answers, we provided an option for participants not wanting to give any details (i.e., "I don't want to state" or "I don't know"). Participants could drop out at any time. Participants had to consent to take part.

3.7 Limitations

We aimed for a representative sample of the German population. Unfortunately, some groups are over- while others are underrepresented. Our sample lacks people of older age, people with lower education, and those with high income. Qualtrics, who acquired the sample for us, stated that this is very common in online surveys. As with every survey study, we have to take into consideration that the data is self-reported. In this study, we additionally asked participants about their future behavior, which is even more prone to uncertainty. Many possible properties of the app have consequences that are not easy to estimate. We cannot assume that participants understood and thought of the consequences, especially considering many participants did not understand how the app worked in detail.

4 Results

In this section, we present the results of the survey. We describe our participants, the accuracy of their knowledge about the CWA, and what sources they consulted. Following that, we describe the participants' intention to use the app and how demographic factors and beliefs about the app explain this decision. Last, we describe how different potential properties, such as additional features, influence the willingness to install the app.

To avoid confusing and overly complicated figures, we assigned short identifiers to each question, which can be seen in Table 6 and Table 7.

4.1 Demographics

Table 1 presents the demographics of the final 744 participants and Table 8 gives an overview of how COVID-19 impacted them.

Since we conducted the study at an early stage of the pandemic, few participants had fallen ill with COVID-19 themselves or had somebody close to them fallen sick. 31.1% count themselves as being a member of the high-risk group. This may seem high but matches estimations in Germany [30]. Around half reported that the pandemic did not influence their work situation (52.7%). 24.5% work from home, and 13.6% reported working in short-time. 74.9% said they did not have specialized tech skills. According to the "Sonntagsfrage" [42],[4] our sample includes 20.7% fewer participants who would have voted for the CDU/CSU[5] at the time the survey was conducted, but 6.6% more participants who would

[4] Regular opinion research in Germany, asking, "Which party would you vote for if federal elections were held this Sunday?"

[5] The Christian Democratic Union of Germany / Christian Social Union in Bavaria

vote for The Greens. All other parties are close to the percentages of the "Sonntagsfrage". We hypothesized the party preference might be an indicator of the attitude towards the app as at least one party publicly criticized the app [33].

4.2 Knowledge

We asked participants to select what they believe are correct statements about the app (Q8-10). As the app was not released when the survey was conducted, answers were not based on experience with the app. However, a press release had been publicized that gave information about the app [27], such as that it would use Bluetooth-Low-Energy, that its primary purpose would be to warn users who had been in close contact with infected people, and that users would not learn who of their contacts reported an infection. At the same time, much misinformation about the app, its purpose, and technical details were spread as well [5].

This section describes the sources participants used and presents participants' beliefs about the to-be-released CWA.

Sources We asked the participants whether and where they heard about the planned corona app (Q7). Figure 3 shows the frequency of how often the participants reported a source. Please note that as participants could report more than one source, the percentages do not add up to 100%. Few but a non-negligible amount of participants (11.7%) reported to never have heard of the app. This leaves 657 participants who were at least somehow aware of the app.

More than half of the participants (54.7%) reported that they received information about the app from public broadcasters. The second most common marked source was social media (29.6%), such as Twitter or Facebook.[6] Scientific publications were used by 7.9% to get information about the CWA.

Correctness of assumptions The following paragraph gives an overview of the participants' assumptions about the CWA (Q8-10). It should be noted that we only included participants who previously reported that they already heard about the app (n=657).

Figure 4 depicts the correct statements for the app that was shortly released after the survey was conducted and shows how many participants marked those to be true. Figure 5 shows all statements that are false for the released app. We classified participants who marked any of the false statements as correct as having "False Beliefs".

59.5% of the participants knew about the app's basic functionality, i.e., that it would warn its users when they had been in contact with another user who later tested positive (`(OTH) INFORMS IF CONTACTED INFECTED` and `(SLF) INFORMS MY CONTACTS`). Around half of the participants knew about the detailed flow that a lab has to confirm the infection before it can be registered in the app (`(SLF) DATA TRANSMISSION ONLY AFTER CONFIRMATION`) to prevent misuse of the app and many false warnings.

However, the app's technical basis was less known: Only 29.8% of the participants who reported to have heard about the app knew that the app would share temporary IDs and timestamps, and 43.5% were aware the app would use Bluetooth. At the same time, 54.6% of the participants thought that the app would use location services, and 24.7% believed the app would use Bluetooth and location services in combination. Although Bluetooth is not a technique developed for position finding, it is, next to GPS, listed as a "location service" in some circumstances [3]. We assume that only participants who marked Bluetooth and location service could have been aware of this detail. 30.0% did not think the app would use Bluetooth but checked location services.

A common misconception (57.5%) was that the app would warn users if an infected person is in their vicinity.

9.89 % of the participants knew all the information that was included in the official press release about the app (`(GEN) SHARES TEMPORARY IDS`, `(GEN) DETECTS NEARBY USERS`, `(GEN) USES BLUETOOTH`, `(GEN) FIGHTS DISEASE SPREAD`, `(OTH) INFORMS IF CONTACTED INFECTED`, `(SLF) INFORMS MY CONTACTS` [27]).

On average, the participants correctly recognized around half of the eight aspects that are true for the app ($median = 4, mean = 4.26, std = 2.05$), but none was known to everyone. Only five participants marked all correct attributes as such and did not believe any incorrect statement.

We asked for the classification of two statements (`(GEN) RESTRICTS BASIC RIGHTS`, `(GEN) THREATS PRIVACY`) that cannot be classified as correct or incorrect but are based on personal sentiments. We saw that participants were worried about their privacy in combination with the app (27.4%) and their basic rights (20.1%). 14.9% stated both in combination.

Misconceptions and lack of information After asking participants how likely they will use the app (Q12), we asked for the primary reason for their installation intention (Q13) in free text form. As we saw many false beliefs, we coded the answers according to underlying misconceptions. We saw statements that were incorrect concerning the app's functionality and its data usage.

Some statements we observed were incorrect but might be correct with the further context of the answer. Participants who (probably) wanted to use the app, for example, stated: *"My safety"*, *"To protect myself"* or *"I want to stay healthy"*. Since the app cannot protect its users directly (users have already been exposed to infected people before they are warned) but only indirectly (the more people download the app, the more people might be influenced and will also download it, leading to better protection of all of its users), these answers indicate a misunderstanding of what the app can do for indi-

[6] Following a statistic from Statista, 65% of the citizens use social media in Germany in general [2].

Gender	**Female**	55.9	**Male**	43.3	**Other**	0.8		
Age	**18-24**	24.2 (9.2)	**25-34**	14.3 (15.3)	**35-49**	23.9 (23.9)	**50-64**	26.8 (26.4)
	65+	10.9 (25.1)						
Education	**ISCED 0-2**	5.9 (16.5)	**ISCED 3-4**	48.9 (58.1)	**ISCED 5-8**	40.5 (25.4)	**Not disclosed**	2.2
Household Income	**<= 1300€**	18.3 (16)	**1300-1700€**	13.4 (8)	**1700-2600€**	20.8 (20)	**2600-3600€**	16.7 (18)
	3600-5000€	14.0 (17)	**>5000€**	6.2 (20)	**Not disclosed**	10.6		
Work Status	**School student**	4.7	**Univ./col. student**	9.8	**Employee**	48.9	**Civil servant**	2.0
	Self-employed	4.6	**Freelancer**	2.2	**Unemployed**	7.5	**Retiree**	16.9
	Not disclosed	3.4						
IT-Knowledge	**Yes**	20.9	**No**	74.9	**Not disclosed**	4.2		
Smartphone OS	**Android**	71.5	**iOS**	26.1	**Other**	2.4		
Political Affiliation	**The Greens**	21.6	**CDU/CSU**	19.4	**SPD**	11.7	**FDP**	6.7
	AfD	6.1	**The Left**	10.1	**Others**	24.5		
Federal State	**BW**	12.9 (13.1)	**BY**	12.1 (15.6)	**BE**	7.0 (4.3)	**BB**	2.4 (3.1)
	HB	1.5 (0.8)	**HH**	3.6 (2.2)	**HE**	6.9 (7.5)	**MV**	2.4 (2.0)
	NI	6.7 (9.6)	**NW**	22.9 (21.6)	**RP**	4.6 (4.9)	**SL**	1.5 (1.2)
	SN	5.0 (5.0)	**ST**	4.7 (2.8)	**SH**	3.5 (3.5)	**TH**	2.4 (2.7)

Table 1: Participants' demographics (N= 744), in percentages. Numbers in brackets = the targeted distribution [17, 29].

vidual users.

Other answers were incorrect beyond doubt. One participant, for example, thought they would be able to see the number of current infections: *"To follow the spread of the pandemic"* (Probably will use the app).[7]

As already seen in Figure 5, participants believed the app would inform its users if infected people are close. This argument was used both as a positive as well as a negative reason to use the app. One participant probably wanted to use the app and argued: *"So I can see who is infected nearby to keep a larger distance to them and protect myself and fellow people."*. Another one did not want to use the app and wrote: *"The determination of the location is too inaccurate. It might happen that other people see me as infected, even though it is somebody else. I have concerns that this might lead to public hostilities or bullying."*

Participants also misunderstood what data will be used and shared: *"I don't want the government to know where I am in each and every second - especially as three other companies are involved as well"*[8] (Definitely will not use it) and *"I don't want the government to have all my numbers and names"* (Definitely will not use it).

Additional to the location misconception, we observed a participant who believed it would be necessary at all times to have access to the internet: *"I don't know how it works but if I need internet you can already forget about it, as I don't have mobile data."*. Anecdotally the participant was not able to correctly answer that the app will use Bluetooth.

Participants indicated that they are confused by the amount of (different) information: *"I don't have any trust. With all the news, I don't know what to believe anymore!!!"* (Probably not use the app). One participant, who failed the comprehension question and was undecided about the app, said: *"Everybody says the opposite of the others. Many say you lose your privacy."*

[7]This is, in fact, possible since version 1.11 which was released at the end of January 2021 [37].

[8]It is not fully clear who the participants refers to. Telekom and SAP developed the app. Two research institutes advised. The RKI is publisher [13]

Following these answers and the data reported previously in this section, we conclude that many participants did not wholly understand the apps' functionality and thus assume a misconception in who will be protected by the app, what data it collects, and with whom the data will be shared.

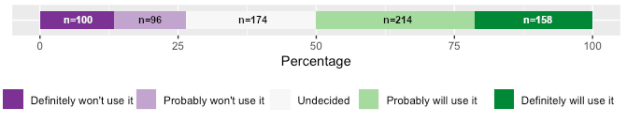


Figure 1: Reported intention to use the app.

4.3 Intention to use

Figure 1 shows the usage intention of all 744 participants. When looking at those participants who were very certain in what they will do, more participants indicated to definitely install the app (Def-Yes, 21.2 %) than to definitely not install it (Def-No, 13.4 %). Almost a third reported they will probably use the CWA (Prob-Yes, 28.8 %) compared to 12.9 % who reported to probably not use it (Prob-No). 23.4 % were still undecided (Undecided) about the installation. As of May 28, 2021 the reported download number of the CWA is 28 millions [22]. That estimates to around 46% of smartphone users in Germany [32]. This estimate does not take into account that that the same person could download the app onto multiple devices.

In the following, we report indications for reasons of the installation intention. For this, we selected an ordered logit regression with a model selection process via best AIC and BIC (c.f. Table 3). In the following paragraphs, we focus the report only on the statistically significant values.

Trust in Government Both trust and distrust of the government correlate heavily with app usage intention. The log odds for both "Somewhat distrust" and "Fully distrust" are proportionally negative compared to the neutral baseline (Log

Odds $= -0.56$ and -1.12 respectively). Contrarily, log odds for both "'Somewhat trust" and "Fully trust" are positive compared to the baseline (Log Odds $= 0.81$ and 1.88 respectively).

Worries Of the worries about future health, economy, and social life, only the health scale was included in the final model. All scale points of this scale are significant and show proportional positive log odds compared to the baseline of "No worries about health" (Log Odds in order of rising concern: 0.53, 0.76, and 1.21). This hints at a positive correlation between future health concerns and app usage intention.

Correlation with beliefs As previously reported, we identified many misconceptions. One of them (`(SLF) GOVERNMENT SEES QUARANTINE VIOLATION`) has a negative impact on the installation intention. Two other attributes that also negatively correlate with it are attributes that can neither be classified as correct or incorrect but are based on personal sentiment: `(GEN) THREATS PRIVACY` (Log Odds $= -1.33$) and `(GEN) RESTRICTS BASIC RIGHTS` (Log Odds $= -1.32$). `(GEN) USES LOCATION SERVICES` likely is an overestimation of the functionality and correlates positively with the intent to install. Another positive correlating attribute is `(GEN) FIGHTS DISEASE SPREAD` (Log Odds $= 0.55$). Its correctness is hard to measure, as there is no central entity that keeps records of how many people were warned by the app and thus ultimately prevented the spread of COVID-19.

Demographics & Personal Experiences We also were interested in which demographic factors and personal experiences with COVID-19 influence participants' decision to use the CWA. We found a statistically significant effect for "Not knowing" whether oneself or someone close was infected by COVID-19. There were negative log odds compared to the baseline of not being infected (Log Odds $= -0.84$). This could be due to a "Don't care" (instead of "Don't know") effect.

4.4 Potential Properties

As mentioned in Section 3, we presented the participants different hypothetical statements and consequences of the app (potential properties), asking whether and how that would influence their decision to use it (Q14). Table 7 in the Appendix shows whether these properties apply to the app as it was described pre-release or not and whether they describe a central or decentral property. We were particularly interested in seeing whether the centralized versus decentralized debate, in which computer scientists and privacy advocates were dominating, was reflected in the broader population's opinions. In the following, we highlight whether the properties belong to the centralized (C) or decentralized (D) approach or if they are independent of the apps' architecture and could be applied for both approaches (B).

Figure 2 shows all potential properties and the distribution of how they would influence the participants. It can be seen that no property is rated exclusively positively or negatively.

However, some have a clear negative tendency (i.e., `(PP) HACKERS KNOW INFECTION STATUS`, `(PP) UNNECESSARY QUARANTINE DUE TO FALSE POSITIVE WARNING`), or a clear positive tendency (`(PP) WARNS ME IF EXPOSED TO COVID`, `(PP) HELPS RKI ASSESS SITUATION`).

Usage intention All potential properties were rated from "Definitely would use it" to "Definitely would not use it". The answers of the participant differ visibly based on the previously stated general usage intention of the app as it was going to be released, i.e., participants who stated that they would want to install the app were more positive about all the potential properties than those who stated that they did not want to use the app and vice versa. We tested this observation with Kruskal-Wallis tests. The results show medium to large effects for all 23 potential properties [50]. This means that per property, there is at least one group that differs from the others in their rating. To find out more, we ran pairwise Wilcoxon rank-sum tests and corrected the p-values with a Bonferroni correction.[9] The poles ("Definitely will use the app" and "Definitely will not use the app") of the installation intention differ from all other groups for each property. Most but not all of the other group comparisons also show a statistically significant difference.

To assess the impact of each property, we report for each group whether the given answer suggests a positive, negative, or no change for the previously stated general intent to use the app. To clarify, if a participant stated that they wanted to install the to-be-released app, then any potential property which was rated "Definitely would use it", "Probably would be willing to use it" and "No influence on my willingness" would lead to no change in their intention and we summarize that as: "No change". However, for the same group, a property rated as either "Definitely would NOT use it" or "Probably would NOT be willing to use it" could lead to a negative effect on the previously positive attitude. We rated these properties as "negative change". The same goes for participants whose general usage intention was negative. Any negative properties would lead to "no change" while a positive property might lead to a "positive change". Participants who stated they were undecided could be swayed in either direction, so only properties rated with "No influence on my willingness" were rated with "no change", and the other received either a positive or negative rating.

We can see large differences between the usage intention groups (cf. Figure 6), especially when looking at the poles of the intention: participants who reported to definitely not use the app (Def-No) (Figure 6a) are seldom really positive

[9] The effect sizes can be seen in the extended version of the paper: https://net.cs.uni-bonn.de/fileadmin/ag/smith/publications/2021_SOUPS_-_CWA.pdf.

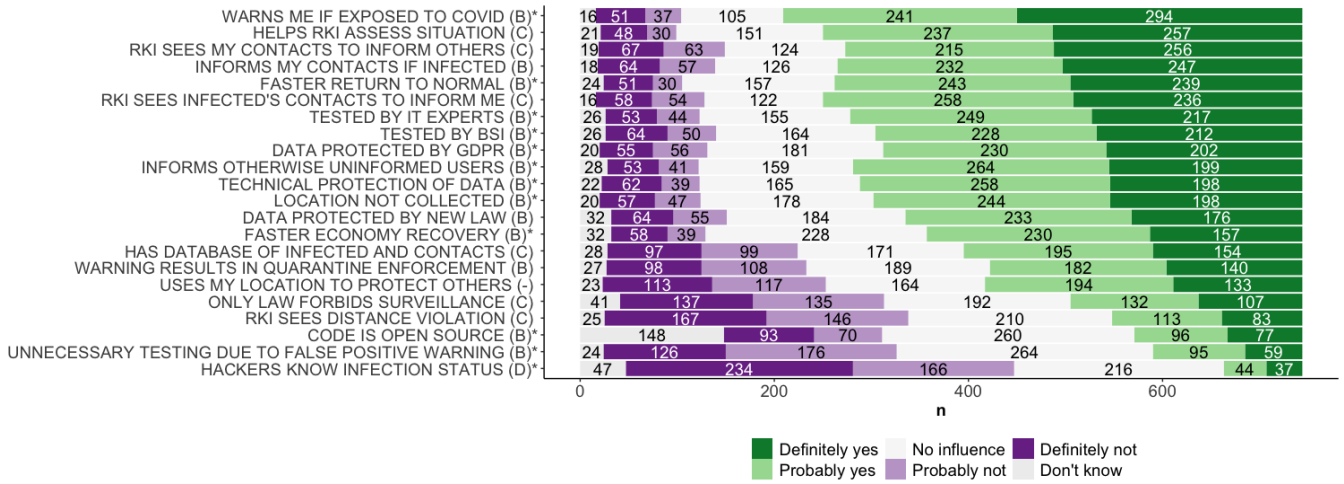


Figure 2: All presented potential properties and the distribution of the ratings of how these would influence the usage intention. * indicates that they apply to the real app. D = decentralized, C = centralized, B = both, "-" = not included in either app design

about any property. In contrast, participants who reported to definitely use the app (Def-Yes) (Figure 6e) are seldom very negative about any property.[10] While we can make no causal claims, the polarisation is noteworthy.

To better assess the individual effects of the different potential properties, we built an ordinal regression model based on a combined score of app usage intention and changes in intention due to these properties (cf. Table 4). Care needs to be taken when interpreting the regression model. Its intention is to highlight the direction of change as described above. However, since both the dependent and independent variables are non-equidistant and contain very strong poles (definitively use/definitively not use), the log odds should probably be seen as an upper bound of the change and needs to be used with care.

Twelve of the potential properties apply to the to-be-released app. Nine of those have a positive effect on usage intention, e.g., `(PP) WARNS ME IF EXPOSED TO COVID` (Log Odds = 1.52) and `(PP) INFORMS OTHERWISE UNINFORMED USERS` (Log Odds = 1.01). Both concern the fact that the app would notify users if they could have been at risk of contracting COVID-19. This was the main feature of the app as communicated to the population. Additionally, the intention to install the app increased if it would help returning to a pre-COVID-19 situation: `(PP) FASTER RETURN TO NORMAL` (Log Odds = 1.17) and `(PP) FASTER ECONOMY RECOVERY` (Log Odds = 0.81).

Two properties that apply to the app impacted the participants negatively: `(PP) HACKERS KNOW INFECTION STATUS` (Log Odds = −1.48) and `(PP) UNNECESSARY TESTING DUE TO FALSE POSITIVE WARNING` (Log Odds = −0.87). The potential of being exposed by a hacker exists [47], but there are methods to mitigate this threat [20]. The risk for unnecessary testing applies to the app, but this could happen without the app and in both the central and decentral approaches.

Eleven potential properties do not apply to the app, of which five have a statistically significant positive influence on the app usage Three of them belong to the centralized approach and offer the Robert Koch-Institute (RKI) additional insights: `(PP) HELPS RKI ASSESS SITUATION` (Log Odds = 1.28), `(PP) RKI SEES MY CONTACTS TO INFORM OTHERS` (Log Odds = 1.12) and `(PP) RKI SEES INFECTED'S CONTACTS TO INFORM ME` (Log Odds = 1.20). `(PP) INFORMS MY CONTACTS IF INFECTED` (Log Odds = 1.15) includes the additional feature of warning users automatically if they had been in contact with an infected person. Currently, users have to actively share their positive test results if they want others to be warned [26].

Three potential properties that do not apply to the app had a negative influence on the installation intention. Two of them (`(PP) RKI SEES DISTANCE VIOLATION` (Log Odds = −0.87) and `(PP) ONLY LAW PREVENTS SURVEILLANCE` (Log Odds = −0.53)) open up the possibility of using the app for surveillance and can fall into the centralized approach; one (`(PP) UNNECESSARY QUARANTINE DUE TO FALSE POSITIVE WARNING` (Log Odds = −1.34)) could be seen as a clear disadvantage for the individual user.

[10] High resolution versions of the figures can be found in the extended version, see Footnote. 9

Trust in different entities Some potential properties are connected to measures taken that should build trust regarding the CWA, regardless of the apps' design choices. These measures included different levels of (data) protection by law, experts testing the app, and the possibility to access the code itself.

As can be seen in Table 4, the idea to protect the data by a new law (`(PP) DATA PROTECTED BY NEW LAW` (Log Odds = 0.73)) as well as the existing protection by the GDPR (`(PP) DATA PROTECTED BY GDPR` (Log Odds = 0.88)) had a positive influence on the intention to use the CWA. Additionally, the technical protection of the data positively influenced the participants (`(PP) TECHNICAL PROTECTION OF DATA` (Log Odds = 1.00)). However, the participants did not seem to like the idea that the government would only be hindered by law to misuse the data for surveillance (`(PP) ONLY LAW PREVENTS SURVEILLANCE` (Log Odds = −0.53)).

It was also rated positively if the CWA would be tested by the German Federal Office for Information Security (BSI) (`(PP) TESTED BY BSI` (Log Odds = 0.9)) and experts (`(PP) TESTED BY IT EXPERTS` (Log Odds = 1.12)).

Interestingly unlike the expert discussion would have suggested, `(PP) CODE IS OPEN SOURCE` did not have a positive effect.

The influence of this property is not statistically significant and it received the most "I don't know" answers compared to all other properties. Even though the terminology "Open Source" is also used in Germany and was communicated in this way in the press release [38] in order to create transparency and trust, we believe many participants lacked an understanding of "Open Source". It thus does not yet seem to have the positive image the technical community would like it to have.

Perception of location services `(PP) LOCATION NOT COLLECTED` had a statistically significant positive influence on the installation intention (Log Odds = 1.10). `(PP) USES MY LOCATION TO PROTECT OTHERS` did not have a significant influence. At the beginning of the survey, we asked participants whether they believed the CWA would use location services. As a reminder: using the users' position was neither the case for the CWA nor was it communicated at any point. However, as mentioned in Section 4.2, the survey question asked about "location services", and Bluetooth may be known as such; therefore, participants could have interpreted it this way. We looked at whether the aspect of location services would make a difference for the installation intention. We compared the participants who a) thought that the CWA would use location services but not Bluetooth to b) those who did not believe the CWA uses location services regarding their general usage intention. 49.8 % of 197 vs 49.7 % of 298. We then also checked whether participants rated the potential properties more positive if they previously indicated that the CWA would use location services. Table 2 shows the percentages of participants who were positively influenced by the property (i.e., answered "Probably would install it" or "Definitely would install it"), split by the general usage intention. As can be seen, the belief that the CWA uses location data did not positively affect the participants' sentiment when being asked how not enabling the government to see their current location would influence them.

5 Discussion

In the following section, we discuss our results, connect them with previous work, and propose directions for future research.

5.1 Participant Beliefs

The majority of the participants knew something about the CWA: Only 5.0% were not able to mark any of the correct app features as true, and the basic idea behind the CWA (that it would warn users with a risk of infection) was known by 59.5% (see Figure 4).

Bluetooth and Location Services We saw a lot of missing information. The technical details that the CWA would use Bluetooth were only known by 43.5%. Interestingly, 30.0% of the participants with some knowledge thought the CWA would use location services but not Bluetooth. . While this topic was discussed quite extensively in the media [10, 40], many people did not seem to think that the CWA would do tracing without GPS or the like. We also hypothesize that many who caught the term "Bluetooth" in the debate did not eliminate GPS from their mental model of the CWA. Another element that could get mixed up with information about the CWA might be the use of cellular network data to measure changes in mobility at the population level, as introduced earlier in the year [34]. Interestingly, we did not see any correlation between the assumption that the CWA uses location services and the usage intention.

Infected Persons Nearby 57.5 % of the participants believed the CWA would warn its users if an infected person is nearby. This was also found by Thomas et al. [66], who studied participants' knowledge regarding the already released Australian app and who found 57.4% of their participants believed this. It was also the most common misconception found by Williams et al. [72] (conducted in the UK). This belief seems very common, even if it was never planned nor (to our knowledge) communicated through official channels that the CWA would be able to warn users of infected persons in their vicinity directly. We are unaware of work that provides insight into why people assume this to be true. However, we hypothesize that many people mixed the two possible app features of being warned afterward and being alerted in real-time. With an incorrect understanding of how contacts are captured and in which cases the infection status is sent or downloaded,

the belief that the CWA could provide real-time warnings is not too far-fetched. Future research should investigate if such vital differences can be communicated, maybe even without going into technical details. It should be noted that this is an overestimation of the CWA's functionality and could lead to incautious behavior based on a false sense of security.

Privacy Concerns 27% of the participants believed the CWA would restrict their basic rights or threats their privacy. These beliefs had significant negative influences on usage intention. Related work found that one of the reasons participants did not want to use an app was because they feared data misuse or surveillance [59, 68]. Some even thought they would receive the names of infected persons [73]. Since the German app (CWA) follows the decentralized approach, only very little data is sent to a central server. While privacy concerns may be valid, we believe many participants did not follow the discussion enough to understand that data storage criticism only concerned the centralized approach. The decentralized app, which was being implemented, stored very little data centrally. We hypothesize they project the worries around the centralized app onto the decentralized one, even if not all concerns are plausible for this approach. Future research is needed to investigate how old mental models can be updated when the underlying system changes and what influences privacy perception. Although, usage intention does not seem to be driven by knowledge about technical details.

Usage Intention We looked at participants' knowledge and beliefs and how they are connected to participants' intentions to use the CWA. Only two attributes to which a correctness value can be assigned had a statistically significant impact on the participants' willingness to use the CWA. The misconception `(OTH) GOVERNMENT SEES QUARANTINE VIOLATION` had a negative impact, the correct attribute `(SLF) INFORMS MY CONTACTS` a positive one. The belief that one's privacy or basic rights were in danger lowered the willingness significantly. It increased if participants thought the CWA would help fight the spread. These assumptions do not reflect knowledge about the app but are based on personal estimations.

As discussed in Section 4.2, `(GEN) USES LOCATION SERVICES` is technically correct in some cases. If participants marked this attribute to be true for the CWA, they were significantly more willing to install the app. Even though the absence of location services as a potential property had a positive influence on using the CWA, participants did not value this absence with a higher usage intention even when previously thinking this would be the case. For this, we have two possible plausible explanations: a) people do not care about location service usage or b) other factors override concerns, e.g., believing in the necessity of the CWA.

Both hypotheses are valuable input for the HCI-community and should be further investigated.

Depending on this, it should be evaluated whether conjoint studies are reliable methods to measure possible acceptance in this domain and how the complexity of reality can be included (i.e., incomplete information or consequential thinking). It seems essential to know the participants' attitudes to the real objective of interest (in our case, the tracing app).

Whether to install the CWA or not seems primarily based on the sentiment of trust and the expectancy of a positive effect. This shows that it is important not only to develop trustworthy technologies but also to communicate their trustworthiness and effectiveness successfully. Technical measures aimed at creating trust do not automatically result in such (e.g., as seen for the CWA's open source property).

5.2 Demographic Factors

A study by Utz et al. [68] was conducted at the same time as this study in Germany and can thus be used to compare the results directly. While the authors conducted an experiment about hypothetical apps and how a tracing app could or should be built, we asked about an app that had been officially announced with a detailed description of features and was near launch. We can confirm part of their findings: We found a positive influence on the willingness to use a corona app a) if the opinion on state government was favorable, b) if participants were concerned about their health, and c) return to a normal life are possible due to the app. Participants with privacy concerns were less likely to use the app, which we can also confirm.

5.3 Never ever or no matter what

Like in our study, other researchers identified participants who did not like any app, regardless of its design choices [58, 62, 68]. We can confirm this finding. Participants within the Def-No group were mostly negative about any of the presented potential properties. 7.1% did not rate a single presented potential property as a positive change. This is similar to the reported 15-21% by Utz et al. [68]. We also saw the exact opposite: Participants belonging to the Def-Yes group rated every single theoretical additional aspect more positively than all other groups. For all potential properties, participants from the installation intention poles (Def-Yes and Def-No) give statistically significantly different answers compared to all other groups.

5.4 Centralized versus Decentralized

Large parts of the discussion around corona tracing apps concerned the technical approach and whether encounters between app users should be stored on a central server or the user's phone. Both approaches come with their advantages and disadvantages. For instance, the centralized app could

give the RKI[11] better insights into how people get infected. Since the central database would be in charge of selecting which users need to be warned, the RKI could see how many people are warned per positive case. Since the risk is computed on the users' devices in the decentralized app, the RKI does not know how many people receive warnings. In the centralized app, it would also have been possible to track how many other positive cases come out of any case, potentially giving more insights into how the virus spreads. On the other hand, the decentralized approach does not facilitate getting an overview but is also not in danger of being extended and misused for surveillance. In general, the centralized app, as it had been planned, offered more insights to healthcare professionals but bore a higher risk of compromise and misuse. However, it is worth noting that the decentralized approach relies on making anonymized infection information public on a central server. This opens the system up for local deanonymization attacks. Suppose an attacker can capture the ephemeral BT-IDs from a target and thus tying those IDs to that target. In that case, they can then monitor the system and see whether they report themselves as positive or not. The German app was based on the decentralized approach (see Section 2) due to public pressure to chose a more privacy-preserving approach. So in the context of this study, we were especially interested in how participants rate the possible benefits and dangers of a centralized app and to see if the debate led by researchers and privacy advocates well represented the feeling of the general public. We included 6 potential properties (Q14) in the survey that were connected to the centralized approach (Table 7). central: All in all, we saw a mix of sentiments. Three central potential properties ((PP) RKI SEES INFECTED'S CONTACTS TO INFORM ME, (PP) RKI SEES MY CONTACTS TO INFORM OTHERS, (PP) HELPS RKI ASSESS SITUATION) had a statistically significant positive influence on the intention to install. All three concern individual or societal benefits. Two other central properties ((PP) RKI SEES DISTANCE VIOLATION, (PP) ONLY LAW PREVENTS SURVEILLANCE) impacted the intention to install negatively. Both focus on the disadvantages of the centralized approach and do not have any clear advantage for the individual user.

The decentral property (PP) HACKERS KNOW INFECTION STATUS impacted the participants in a negative way. While this risk is limited to local attackers, and there are methods to mitigate this threat [20], it is something that our participants did not like. However, it did not feature significantly in the public debate as far as we know and, as such, is unlikely to have had much of an impact.

It seems participants are in general inclined to rate properties of the centralized approach positively while they rate the consequences (in the current technical landscape) that come with it rather low.

In summary, many of our participants had very positive views concerning the increased capabilities the centralized app would have had. This suggests that there could have been more support in the population for a more feature-rich app than academics and privacy advocates acknowledged in the discussion preceding the CWA's publication. Relevant health officials have since stated that the app in its current form is no great support [16], and due to the privacy design, it is hard to evaluate its efficacy. We think it is worth discussing whether a more nuanced discussion about the feature/privacy trade-off would be warranted for the future.

6 Conclusion

We surveyed the usage intention of the CWA in Germany right before its launch. 50% of the participants reported their intent to use the CWA, 26.3% refrained from usage and 23.4% were undecided. This seems reasonably close to the most recent (May 28, 2021) download numbers. To understand their decision, we investigated what beliefs participants had about the CWA. We saw many false beliefs, especially concerning technical details, i.e., 30.0% of the participants thought the CWA would use location services (other than Bluetooth). Actual knowledge about the CWA does not seem to be the primary driver for the decision to use the CWA. Instead, perceived privacy or basic rights intrusions led to a lower intention to use it. As also reported by other researchers, we found a positive effect when people were worried about general health and trusted the government. We also highlight that the general population's views were more diverse and more open to a central entity getting an overview to help fight the pandemic than the public discussion indicated. Based on our results, we recommend future work on a) where the privacy concerns come from, as in our view many of the concern did not match the actual CWA and b) how the perceptions can be aligned with the actual facts of the CWA, as this is necessary to discuss features based on the facts. And c) whether the CWA can be extended in a way that it becomes more useful to the relevant parties, e.g., the public health departments, while at the same time implementing technical countermeasures to prevent the data from being abused.

Acknowledgements

We thank, in alphabethical order: Yomna Abdelrahman, Ruba Ali Mahmoud Abu-Salma, Florian Alt, Zinaida Benenson, Nataliia Bielova, Freya Gassmann, Katharina Krombholz, Mattia Mossano and Melanie Volkamer for their insightful discussions and helpful remarks on the survey. This work was partially funded by the Werner Siemens Foundation.

[11] According to their website, "The Robert Koch Institute (RKI) is the government's central scientific institution [a federal government agency] in the field of biomedicine. It is one of the most important bodies for the safeguarding of public health in Germany." [31]

References

[1] AIC vs. BIC. https://www.methodology.psu.edu/resources/AIC-vs-BIC/. Accessed: June 09, 2021.

[2] Anteil der Nutzer von Social Networks in Europa nach ausgewählten Ländern 2020. https://de.statista.com/statistik/daten/studie/214663/umfrage/nutzung-von-social-networks-in-europa-nach-laendern/. Accessed: February 19, 2021.

[3] Bluetooth low energy. https://developer.android.com/guide/topics/connectivity/bluetooth-le#permissions. Accessed: May 29, 2021.

[4] Bundesinnenminister: Corona-Warn-App erfüllt höchste Ansprüche an Sicherheit und Datenschutz. https://www.bmi.bund.de/SharedDocs/kurzmeldungen/DE/2020/06/vorstellung-corona-warn-app.html. Accessed: February 17, 2021.

[5] Bundesregierung - Mythen und Falschmeldungen - Corona-Warn-App. https://web.archive.org/web/20200616193256/https://www.bundesregierung.de/breg-de/themen/mythen-und-falschmeldungen/corona-app-falschmeldungen-1758136. Accessed: February 02, 2021.

[6] Clickworker. www.clickworker.de/. Accessed: February 04, 2021.

[7] Contact Tracing Joint Statement. https://www.esat.kuleuven.be/cosic/sites/contact-tracing-joint-statement/. Accessed: January 28, 2021.

[8] CoroBlog zur wissenschaftlichen Auswertung der Corona-Datenspende-App. https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/Corona-Datenspende.html. Accessed: January 26, 2021.

[9] Corona-Apps im Überblick – Digitale Informations- und Hilfsangebote. https://www.zusammengegencorona.de/informieren/corona-warn-app/corona-apps-im-ueberblick/. Accessed: May 26, 2021.

[10] Corona-Pandemie: Alles Wissenswerte rund um die Warn-App. https://www.tagesschau.de/inland/faq-corona-tracing-app-101.html. Accessed: February 19, 2021.

[11] Corona-Tracing-App: Offener Brief an Bundeskanzleramt und Gesundheitsminister. https://www.ccc.de/de/updates/2020/corona-tracing-app-offener-brief-an-bundeskanzleramt-und-gesundheitsminister. Accessed: February 23, 2021.

[12] Corona-Tracing: Bundesregierung denkt bei App um. https://www.tagesschau.de/inland/coronavirus-app-107.html. Accessed: January 28, 2021.

[13] Corona-Warn-App: Fragen und Antworten. https://www.bundesregierung.de/breg-de/themen/corona-warn-app/corona-warn-app-faq-1758392. Accessed: February 23, 2021.

[14] Corona-Warn-App: Fragen und Antworten. https://www.verbraucherzentrale.de/wissen/digitale-welt/apps-und-software/coronawarnapp-fragen-und-antworten-zur-deutschen-tracingapp-47466. Accessed: February 17, 2021.

[15] Corona-Warn-App: Github. https://github.com/corona-warn-app. Accessed: February 17, 2021.

[16] Corona-Warn-App: Wenige Infektionen gemeldet (released October 2, 2020). https://www.aerzteblatt.de/archiv/216016/Corona-Warn-App-Wenige-Infektionen-gemeldet. Accessed: February 22, 2021.

[17] Einnahmen und Ausgaben privater Haushalte. https://www.destatis.de/DE/Themen/Gesellschaft-Umwelt/Einkommen-Konsum-Lebensbedingungen/Einkommen-Einnahmen-Ausgaben/_inhalt.html. Accessed: May 29, 2020.

[18] Faktencheck: Nein, die Corona-Warn-App nutzt keine persönlichen Kontaktdaten. https://correctiv.org/faktencheck/2020/06/25/nein-die-corona-warn-app-nutzt-keine-persoenlichen-kontaktdaten/. Accessed: February 17, 2021.

[19] Fünf Monate nach dem Start: Die Tücken der Corona-Warn-App. https://www.tagesschau.de/inland/faq-corona-warn-app-101.html. Accessed: February 17, 2021.

[20] Google Exposure Notification Key Server. https://google.github.io/exposure-notifications-server/server_functional_requirements.html. Accessed: May 28, 2021.

[21] Help speed up contact tracing with TraceTogether. https://www.gov.sg/article/help-speed-up-contact-tracing-with-tracetogether. Accessed: February 09, 2021.

[22] Kennzahlen zur Corona-Warn-App. https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/WarnApp/Archiv_Kennzahlen/Kennzahlen_28052021.pdf?__blob=publicationFile. Accessed: June 1, 2021.

[23] Offener Brief: Geplante Corona-App ist höchst problematisch. http://www.fiff.de/presse/coronaappproblematisch. Accessed: January 28, 2021.

[24] Pan-European Privacy-Preserving Proximity Tracing. https://github.com/pepp-pt/pepp-pt-documentation/blob/master/10-data-protection/PEPP-PT-data-protection-information-security-architecture-Germany.pdf. Accessed: January 22, 2021.

[25] Pläne von Minister Spahn - Kritik an Corona-App wächst. https://www.tagesschau.de/inland/corona-app-spahn-101.html. Accessed: January 28, 2021.

[26] Poster: Positiv getestet? So teilen Sie Ihr Ergebnis über die Corona-Warn-App. https://www.rki.de/DE/Content/InfAZ/N/Neuartiges_Coronavirus/WarnApp/Poster.pdf?__blob=publicationFile. Accessed: June 09, 2021.

[27] Pressemitteilung des Bundesministeriums für Gesundheit, des Bundesministeriums des Innern, für Bau und Heimat und des Bundeskanzleramts zum Projekt "Corona-App" der Bundesregierung. https://www.bundesregierung.de/breg-de/aktuelles/pressemitteilung-des-bundesministeriums-fuer-gesundheit-des-bundesministeriums-des-innern-fuer-bau-und-heimat-und-des-bundeskanzleramts-zum-projekt-corona-app-der-bundesregierung-1747916. Accessed: January 28, 2021.

[28] Privacy notice CWA. https://www.coronawarn.app/assets/documents/cwa-privacy-notice-en.pdf. Accessed: February 23, 2021.

[29] Qualtrics. qualtrics.com. Accessed: January 22, 2021.

[30] Risikogruppen sind überall. https://de.statista.com/infografik/21145/groesse-von-ausgewaehlten-risikogruppen-in-deutschland/. Accessed: February 23, 2021.

[31] RKI - The Institute. https://www.rki.de/EN/Content/Institute/institute_node.html. Accessed: February 15, 2021.

[32] Statistiken zur Smartphone-Nutzung in Deutschland. https://de.statista.com/themen/6137/smartphone-nutzung-in-deutschland/. Accessed: February 23, 2021.

[33] Sylvia Limmer: Corona-Appidemie stoppen! https://www.afd.de/sylvia-limmer-corona-appidemie-stoppen/. Accessed: May 28, 2021.

[34] Süddeutsche Zeitung GmbH. https://sz.de/1.4850094. Accessed: February 25, 2021.

[35] Timeline of ECDC's reponse to COVID-19. https://www.ecdc.europa.eu/en/covid-19/timeline-ecdc-response. Accessed: February 09, 2021.

[36] Transkript: Erklärfilm zur „Corona-Warn-App“. https://www.bundesregierung.de/resource/blob/1726066/1760258/c4247487c176123b13003e6125ead7aa/2020-06-10-corona-warn-app-de-textversion-data.pdf. Accessed: February 17, 2021.

[37] Version 1.11: Corona-Warn-App nun mit Kennzahlen zum Infektionsgeschehen. https://www.coronawarn.app/de/blog/2021-01-28-corona-warn-app-version-1-11/. Accessed: February 15, 2021.

[38] Veröffentlichung der Corona-Warn-App. https://www.bundesregierung.de/breg-de/themen/coronavirus/veroeffentlichung-der-corona-warn-app-1760892. Accessed: February 01, 2021.

[39] "Vorbildlich gelaufen" -Chaos Computer Club lobt deutsche Corona-App. https://www.zdf.de/nachrichten/politik/corona-app-launch-100.html. Accessed: February 17, 2021.

[40] Wann kommt die App, die hilft? https://www.zeit.de/digital/datenschutz/2020-04/corona-app-tracking-handydaten-bluetooth-datenschutz. Accessed: February 19, 2021.

[41] Washington Post-University of Maryland national poll, April 21-26, 2020. https://web.archive.org/web/20200501144955if_/https://www.washingtonpost.com/context/washington-post-university-of-maryland-national-poll-april-21-26-2020/3583b4e9-66be-4ed6-a457-f6630a550ddf/. Accessed: February 10, 2021.

[42] Wenn am nächsten Sonntag Bundestagswahl wäre ... https://www.wahlrecht.de/umfragen/forsa.htm. Accessed: January 27, 2021.

[43] Zwischenbilanz der Corona-Nachverfolgung: Wie erfolgreich funktionieren Corona-Apps in der EU? https://www.treffpunkteuropa.de/zwischenbilanz-der-corona-nachverfolgung-wie-erfolgreich-funktionieren. Accessed: February 23, 2021.

[44] “Decentralised Privacy-Preserving Proximity Tracing. https://github.com/DP-3T/documents/blob/master/DP3T%20White%20Paper.pdf. Accessed: January 22, 2021.

[45] Samuel Altmann, Luke Milsom, Hannah Zillessen, Raffaele Blasone, Frederic Gerdon, Ruben Bach, Frauke Kreuter, Daniele Nosenzo, Séverine Toussaert, and Johannes Abeler. Acceptability of app-based contact tracing for COVID-19: Cross-country survey study. *JMIR mHealth and uHealth*, 8(8):e19857, 2020.

[46] Patrik Bachtiger, Alexander Adamson, Jennifer K. Quint, and Nicholas S. Peters. Belief of having had unconfirmed Covid-19 infection reduces willingness to participate in app-based contact tracing. *NPJ digital medicine*, 3(1):1–7, 2020.

[47] Lars Baumgärtner, Alexandra Dmitrienko, Bernd Freisleben, Alexander Gruler, Jonas Höchst, Joshua Kühlberg, Mira Mezini, Markus Miettinen, Anel Muhamedagic, Thien Duc Nguyen, et al. Mind the gap: Security & privacy risks of contact tracing apps. *arXiv preprint arXiv:2006.05914*, 2020.

[48] Fabian Buder, Anja Dieckmann, Vladimir Manewitsch, Holger Dietrich, Caroline Wiertz, Aneesh Banerjee, Oguz A. Acar, and Adi Ghosh. Adoption Rates for Contact Tracing App Configurations in Germany, 2020.

[49] Kenneth P. Burnham and David R. Anderson. Multimodel inference: understanding AIC and BIC in model selection. *Sociological methods & research*, 33(2):261–304, 2004.

[50] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[51] Jemima A. Frimpong and Stephane Helleringer. Financial Incentives for Downloading COVID–19 Digital Contact Tracing Apps. SocArXiv 9vp7x, Center for Open Science, June 2020.

[52] M. Guillon and P. Kergall. Attitudes and opinions on quarantine and support for a contact-tracing application in France during the COVID-19 outbreak. *Public health*, 188:21–31, 2020.

[53] Eszter Hargittai, Elissa M. Redmiles, Jessica Vitak, and Michael Zimmer. Americans’ willingness to adopt a COVID-19 tracking app. *First Monday*, 2020.

[54] Farkhondeh Hassandoust, Saeed Akhlaghpour, and Allen C. Johnston. Individuals’ privacy concerns and adoption of contact tracing mobile applications in a pandemic: A situational privacy calculus perspective. *Journal of the American Medical Informatics Association*, 2020.

[55] Robert Hinch, Will Probert, Anel Nurtay, Michelle Kendall, Chris Wymant, Matthew Hall, Katrina Lythgoe, Ana Bulas Cruz, Lele Zhao, Andrea Stewart, et al. Effective configurations of a digital contact tracing app: A report to nhsx. *Retrieved July*, 23:2020, 2020.

[56] Kai T. Horstmann, Susanne Buecker, Julia Krasko, Sarah Kritzler, and Sophia Terwiel. Who does or does not use the “Corona-Warn-App” and why? *European Journal of Public Health*, 2020.

[57] Laszlo Horvath, Susan Banducci, and Oliver James. Citizens’ Attitudes to Contact Tracing Apps. *Journal of Experimental Political Science*, pages 1–13, 2020.

[58] Gabriel Kaptchuk, Eszter Hargittai, and Elissa M. Redmiles. How good is good enough for COVID19 apps? The influence of benefits, accuracy, and privacy on willingness to adopt. *arXiv preprint arXiv:2005.04343*, 2020.

[59] Genia Kostka and Sabrina Habich-Sobiegalla. In Times of Crisis: Public Perceptions Towards COVID-19 Contact Tracing Apps in China, Germany and the US. *Germany and the US (September 16, 2020)*, 2020.

[60] Wassili Lasarov. Im Spannungsfeld zwischen Sicherheit und Freiheit. *HMD Praxis der Wirtschaftsinformatik*, pages 1–18, 2020.

[61] Tianshi Li, Camille Cobb, Sagar Baviskar, Yuvraj Agarwal, Beibei Li, Lujo Bauer, Jason I. Hong, et al. What Makes People Install a COVID-19 Contact-Tracing App? Understanding the Influence of App Design and Individual Difference on Contact-Tracing App Adoption Intention. *arXiv preprint arXiv:2012.12415*, 2020.

[62] Tianshi Li, Cori Faklaris, Jennifer King, Yuvraj Agarwal, Laura Dabbish, Jason I. Hong, et al. Decentralized is not risk-free: Understanding public perceptions of privacy-utility trade-offs in COVID-19 contact-tracing apps. *arXiv preprint arXiv:2005.11957*, 2020.

[63] Michael E. O’Callaghan, Jim Buckley, Brian Fitzgerald, Kevin Johnson, John Laffey, Bairbre McNicholas, Bashar Nuseibeh, Derek O’Keeffe, Ian O’Keeffe, Abdul Razzaq, et al. A national survey of attitudes to COVID-19 digital contact tracing in the Republic of Ireland. *Irish Journal of Medical Science*, pages 1–25, 2020.

[64] Elissa M. Redmiles. User Concerns & Tradeoffs in Technology-facilitated COVID-19 Response. *Digital Government: Research and Practice*, 2(1):1–12, 2020.

[65] Lucy Simko, Jack L. Chang, Maggie Jiang, Ryan Calo, Franziska Roesner, and Tadayoshi Kohno. COVID-19 Contact Tracing and Privacy: A Longitudinal Study of Public Opinion. *arXiv preprint arXiv:2012.01553*, 2020.

[66] Rae Thomas, Zoe A. Michaleff, Hannah Greenwood, Eman Abukmail, and Paul Glasziou. Concerns and Misconceptions About the Australian Government’s COVIDSafe App: Cross-Sectional Survey Study. *JMIR public health and surveillance*, 6(4):e23081, 2020.

[67] Simon Trang, Manuel Trenz, Welf H. Weiger, Monideepa Tarafdar, and Christy MK Cheung. One app to trace them all? Examining app specifications for mass acceptance of contact-tracing apps. *European Journal of Information Systems*, 29(4):415–428, 2020.

[68] Christine Utz, Steffen Becker, Theodor Schnitzler, Florian M. Farke, Franziska Herbert, Leonie Schaewitz, Martin Degeling, and Markus Dürmuth. Apps against the spread: Privacy implications and user acceptance of covid-19-related smartphone apps on three continents.

In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[69] Viktor von Wyl, Marc Hoeglinger, Chloe Sieber, Marco Kaufmann, Andre Moser, Miquel Serra-Burriel, Tala Ballouz, Dominik Menges, Anja Frei, and Milo A. Puhan. Drivers of acceptance of COVID-19 proximity tracing apps in Switzerland. *medRxiv*, 2020.

[70] Michel Walrave, Cato Waeterloos, and Koen Ponnet. Ready or Not for Contact Tracing? Investigating the Adoption Intention of COVID-19 Contact-Tracing Technology Using an Extended Unified Theory of Acceptance and Use of Technology Model. *Cyberpsychology, Behavior, and Social Networking*, 2020.

[71] Caroline Wiertz, Aneesh Banerjee, Oguz A. Acar, and Adi Ghosh. Predicted Adoption Rates of Contact Tracing App Configurations-Insights from a choice-based conjoint study with a representative sample of the UK population. *Available at SSRN 3589199*, 2020.

[72] Simon N. Williams, Christopher J. Armitage, Tova Tampe, and Kimberly Dienes. Public attitudes towards COVID-19 contact tracing apps: A UK-based focus group study. *Health Expectations*, 2020.

[73] Baobao Zhang, Sarah Kreps, Nina McMurry, and R. Miles McCain. Americans' perceptions of privacy and surveillance in the COVID-19 pandemic. *Plos one*, 15(12):e0242652, 2020.

A Survey

Screening Questions

- Q1 What is your age?
  [Free Text]
- Q2 In which federal state do you live?
- Q3 Do you use a smartphone?
  [Yes, an Android / Yes, an IPhone / Yes, another smartphone / Yes, but I don't know which / No / I don't want to state]
- Q4 What is your netto household income?
  [<= 1300 / 1300-1700€ / 1700-2600€ / 2600-3600€ / 3600-5000€ / > 5000 / I don't want to state]
- Q5 What is the number of individuals living in your household?
  [1 / 2 / 3 / 4 or more / I don't want to state]
- Q6 What is the highest-level vocational qualification you hold?
  [Completed apprenticeship / Other; Vocational qualification: / University degree / Master or Technician certification or equivalent technical school diploma / Vocational school diploma / Technical school diploma / No vocational qualification / Technical college degree (or engineering school diploma) / I don't want to state / Abitur (German university entrance qualification)]

App Description and Media Sources

The COVID-19 coronavirus pandemic is a worldwide problem. The Corona warning app for Germany is one of the measures planned to assist health authorities in tracing and containing infection, being developed by SAP to run on Deutsche Telekom infrastructure. The Robert Koch Institute (RKI) will publish the app when it is ready. It is also referred to as the 'Corona app', 'COVID app' or 'contact tracing app'.

- Q7 Have you heard of the plans for this app? If 'yes', please select where you heard about the app. Multiple selections possible.
  [Public broadcasters (ARD, ZDF, WDR, etc.) / Non-public TV (Pro7, Vox, N24, etc.) / Scientific publications / Newspapers, journals, magazines, etc. / Family member / Official government/state agency websites (Robert Koch Institute, Federal Government, etc.) / Other websites: / I have not heard about this app / Friends / Social media (Twitter, Facebook, YouTube, TikTok, etc.) / Work colleagues/associates / Don't know/I don't want to state / Official Corona Warning App website

Knowledge

- Q8 Which of the below statements do you think will apply regarding the app? (please check all that apply.)
  - The app uses Bluetooth.
  - Through the app I can donate health data to the Robert Koch Institute for research purposes.
  - The app determines when other smartphones are nearby that are also using the app.
  - The app shares temporary IDs and timestamps.
  - The app enables the government to see my current location.
  - The app enables the government to see if people are not keeping a safe distance from others.
  - Usage of the app will be mandatory.
  - The app shares the names and phone numbers of my contacts with the government.
  - The app infringes my basic rights.
  - The app can be used to demonstrate to others that I am not currently COVID-19 positive.
  - The app facilitates decision-making on who should be tested for COVID-19.
  - The app shares fitness data.
  - The app can help fight the spread of the COVID-19 virus.
  - The app uses location services (like GPS).
  - The app shares a profile of my movement.
  - None of the above applies.
  - Don't know
  - The app undermines my privacy.
- Q9 What statements do you think apply regarding the app when other users are COVID-19 positive? (please check all that apply.)
  - The app enables the government to see if someone is not complying with quarantine orders.
  - The app notifies me if I have had contact with an individual who later tested positive for COVID-19.
  - The app notifies me when an infected person is located nearby.
  - None of the above applies.
  - Don't know
- Q10 What statements do you think apply regarding the app when you yourself are COVID-19 positive? (please check all that apply.)
  - The app informs other app users who have been close to me that they may have contracted the virus.
  - The app sends data continuously to the RKI.
  - A physician or the public health authority has to confirm my positive COVID-19 test result before the app sends data to the RKI.
  - The app enables the government to see if I am not complying with quarantine orders.
  - None of the above applies.
  - Don't know

App Description and Comprehension

A brief introduction is provided below on the planned capabilities of the contact tracing app. The federal government intends to introduce a smartphone app to trace COVID-19 transmission in the near future. The app is to be very user-friendly and its usage voluntary. The app is designed to ensure that virus transmission is detected more quickly. This allows taking targeted containment measures.
When in use, the app determines what other users of the app are located near you. The app does this via Bluetooth. The app will alert you if you have been near someone within the past few days who subsequently tested positive for COVID-19. The app then informs you of what you need to do next, such as get tested for COVID-19.

- Q11 How will the described app determine what people have been near me?
  [Bluetooth / Location services (such as GPS) / My phone Contacts list / Don't know]

Install General

In answering the following questions, please imagine that the app described above has already been released. The app is being developed by SAP to run on Deutsche Telekom infrastructure. The Robert Koch Institute (RKI) is in charge of the app and evaluates the data. The exclusive permissible usage of the data is to fight COVID-19.

- Q12 How likely is it that you will use the app?
  [Definitely will use it / Probably will use it / Undecided / Probably will not use it / Definitely will not use it / Response declined / Don't know]
- Q13 What is the primary reason for your answer?
  [Free text]

Potential Properties

Q14 You will now be presented with 24 statements. These statements concern characteristics or things that **could** apply or be true with the app. Please select how these statements, **if true**, would influence your willingness to use the app.
[Definitely would use it / Probably would be willing to use it / No influence on my willingness / Probably would not be willing to use it / Definitely would not use it / Don't know]

- The government would be prevented by law, but not by technical means, from misusing the data for surveillance purposes.
- Using the app would enable the RKI to find out if I am not complying with minimum distancing to other individuals.
- The RKI would have a database with the contact data of infected individuals and the people they have had contact with.
- If I test positive for COVID-19, the app would allow the RKI to see who I had contact with in order to notify those individuals
- The German Federal Office for Information Security (BSI) would verify that the app fulfills data security and data protection requirements.
- Using the app would make possible a speedier return to normal public life.
- Technical measures would be implemented to ensure the data are protected.
- There is a possibility that the app could incorrectly report infection risk, resulting in me having to quarantine unnecessarily
- Using the app would help re-start the economy faster
- If the app notifies me that I may have been infected, I would have be required by law to quarantine.
- The app would notify me if I have been in a situation putting me at risk of contracting COVID-19.
- There is a possibility that the app could incorrectly report infection risk, resulting in me having to get tested unnecessarily
- Independent security experts would verify that the app fulfills data security and data protection requirements.
- The app would use information about my location to more accurately monitor infection risk for others
- Protection of the data would be guaranteed pursuant to a data protection policy and the General Data Protection Regulation.
- The app would not collect any data about my location.
- The app would inform people of infection risk who would not otherwise be contacted by the public health authority.
- Any nearby hackers could find out if I have tested positive for COVID-19.
- This question pertains to attentive completion of the survey. Please select "No influence" as response.
- If somebody near me has tested positive for COVID-19, the app would enable the RKI to see that I have had contact with that individual in order to notify me accordingly.
- Protection of the data would be guaranteed under a new law drafted especially for the app
- If I have tested positive for COVID-19, the app would automatically notify other users of the app who are at risk being exposed through contact with me
- The app would be open-source
- The app would support the RKI to better assess the COVID-19 situation.

It is being discussed whether use of the app should be made mandatory in certain situations where people come in contact in groups, such as patronizing restaurants or utilizing bus or train services, to facilitate targeted monitoring of infection risk. It must be considered however that roughly 20% of the German population would be excluded from using such services due to not having a smartphone.

- Q15 Would you approve or disapprove of such mandatory usage?
  [Approve entirely / Mainly approve / Neither approve nor disapprove / Mainly disapprove / Disapprove entirely / Response declined / Don't know]

Demographics

- Q16 What is your gender?
  [Male / Female / Non-binary / Would like to self-describe: / I don't want to state]
- Q17 What is your work status?
  [School student / University/college student / Employee / Civil servant / Self-employed / Freelancer / Unemployed / Retiree / I don't want to state]
- Q18 Do you have specialized computing skills, such as: system administration, programming, IT security, tech support, power user, etc?
  [Yes / No / I don't want to state]
- Q19 Please indicate your agreement or disagreement with the following: "I generally trust the government to do the right thing."
  [Fully agree / Mostly agree / Neither agree nor disagree / Mostly disagree / Fully disagree / I don't want to state]
- Q20 What party do you have the most affinity with?
  [The Greens / CDU/CSU / SPD / FDP / AfD / The Left / Others/I don't want to state]
- Q21 Currently, how frequently do you have close personal contact with people not from your household?
  [Once a week at most / A few times a week / A few times a day / Several times a day / I don't want to state]

- Q22 How concerned or unconcerned are you about COVID-19 in regard to the following three areas?
  Health, The economy, Society
  [Unconcerned / A bit concerned / Concerned / Very concerned / I don't want to state]
- Q23 Do you fall within a COVID-19 high-risk group?
  [Yes / No / Don't know / I don't want to state]
- Q24 Does someone close to you fall within a COVID-19 high-risk group?
  [Yes / No / Don't know / I don't want to state]
- Q25 Have you or any person close to you fallen ill with Covid-19?
  [Yes / No / Don't know / I don't want to state]
- Q26 Has anyone close to you died of Covid-19?
  [Yes / No / Don't know / I don't want to state]
- Q27 How has the Covid-19 pandemic affected you financially?
  [Positive impact / No impact / Negative impact / Critical impact / I don't want to state]
- Q28 Has the Covid-19 pandemic resulted in you having to look after/care for someone at home?
  [Yes / No / I don't want to state]
- Q29 How has the crisis affected your work?
  [Unaffected / Working from home / Short-time work / Became unemployed / Found employment / I don't want to state]

B Additional Tables and Figures

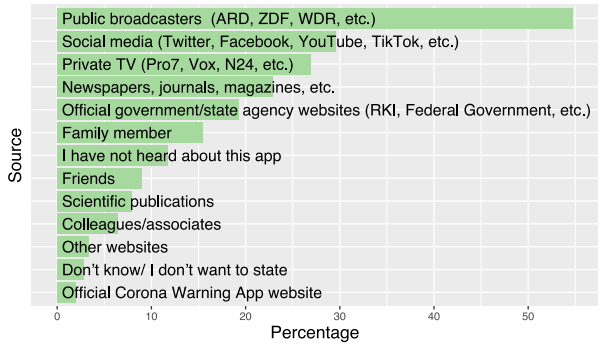


Figure 3: Frequency of reported information sources (n=744).

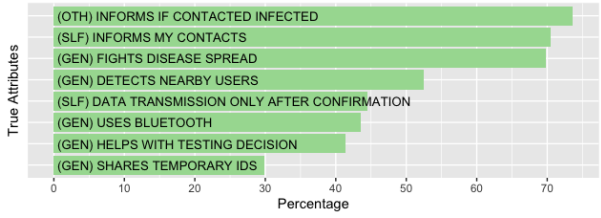


Figure 4: Attributes that are correct for the current app and the percentage of participants who checked the corresponding box. OTH: other is infected, SLF: self infected, GEN: general attribute

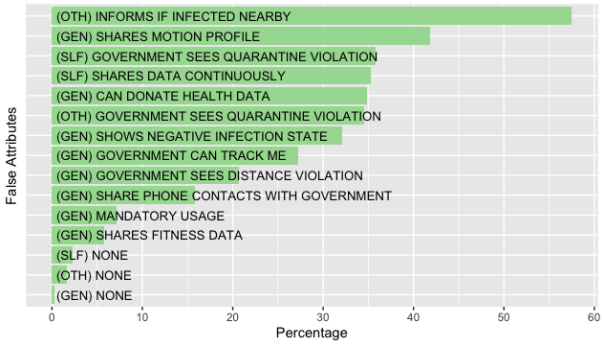


Figure 5: Attributes that are wrong for the current app and the percentage of participants that checked the corresponding box.

Usage intention	LBNB belief?	Positive influence if no location usage
Undecided	Yes (n= 46)	26.1%
	No (n= 108)	28.7%
Prob-No	Yes (n= 27)	3.7%
	No (n= 53)	15.1%
Def-No	Yes (n= 25)	4%
	No (n= 65)	7.7%

Table 2: Percentage of participants who rated the potential property that the app would not collect data about users' position positively based on their general usage intention and whether they believed the app would be working with location data. LBNB = Location service but no Bluetooth.

Factor	Log Odds	C.I.	p-value
Trust in Government (Q19)			
Trust: Fully agree	1.88	[1.26, 2.50]	$< 0.001*$
Trust: Somewhat agree	0.81	[0.41, 1.20]	$< 0.001*$
Trust: Somewhat disagree	−0.56	[−1.08, −0.04]	$0.035*$
Trust: Fully disagree	−1.12	[−1.85, −0.39]	$0.003*$
Beliefs			
(GEN) THREATS PRIVACY	−1,33	[−1.82, −0.84]	$< 0.001*$
(GEN) FIGHTS DISEASE SPREAD	0.55	[0.16, 0.94]	$0.005*$
(GEN) RESTRICTS BASIC RIGHTS	−1.32	[−1.88, −0.77]	$< 0.001*$
(OTH) GOVERNMENT SEES QUARANTINE VIOLATION	−0.70	[−1.08, −0.33]	$< 0.001*$
(SLF) INFORMS MY CONTACTS	0.51	[0.15, 0.88]	$0.006*$
(GEN) MANDATORY USAGE	0.66	[−0.07, 1.39]	0.075
(GEN) USES LOCATION SERVICES	0.42	[0.06, 0.77]	$0.022*$
(SLF) DATA TRANSMISSION ONLY AFTER CONFIRMATION	0.31	[−0.02, 0.65]	0.069
(OTH) INFORMS IF INFECTED NEARBY	−0.28	[−0.61, 0.06]	0.107
Worries (Q22)			
Health: Somewhat worried	0.53	[0.04, 1.02]	$0.036*$
Health: Worried	0.76	[0.24, 1.28]	$0.004*$
Health: Very worried	1.21	[0.63, 1.79]	$< 0.001*$
Media Sources (Q7)			
Media: Off. Homepage	0.99	[−0.14, 2.12]	0.085
Media: Publications	0.62	[0.07, 1.18]	$0.028*$
Media: Public Broadcasters	−0.30	[−0.63, 0.04]	0.082
Personal Experience (Q25)			
Was or knows infected: Yes	−0.06	[−0.63, 0.51]	0.840
Was or knows infected: Don't know	−0.84	[−1.57, −0.11]	$0.024*$
Demographics (Q16, Q18)			
Tech Background	0.24	[−0.14, 0.62]	0.208
Intercepts (App usage intention)			
Definitely not | Probably not	−1.99	[−2.61, −1.37]	$< 0.001*$
Probably not | Undecided	0.35	[0.14, 0.56]	$0.001*$
Undecided | Probably would	0.53	[0.38, 0.68]	$< 0.001*$
Probably would | Definitely would not	0.61	[0.48, 0.74]	$< 0.001*$

Table 3: Results of the final ordered logit regression model correlating factors with app usage intention. "Don't want to answer" answers were omitted. See Section 3.5 and Table 5 for further details.

Factor	Log Odds	C.I.	p-value
(PP) WARNS ME IF EXPOSED TO COVID	1.52	[1.31, 1.73]	$< 0.001*$
(PP) INFORMS MY CONTACTS IF INFECTED	1.15	[0.94, 1.37]	$< 0.001*$
(PP) INFORMS OTHERWISE UNINFORMED USERS	1.01	[0.80, 1.23]	$< 0.001*$
(PP) HELPS RKI ASSESS SITUATION	1.28	[1.07, 1.49]	$< 0.001*$
(PP) FASTER RETURN TO NORMAL	1.17	[0.96, 1.39]	$< 0.001*$
(PP) FASTER ECONOMY RECOVERY	0.81	[0.59, 1.03]	$< 0.001*$
(PP) RKI SEES MY CONTACTS TO INFORM OTHERS	1.12	[0.90, 1.33]	$< 0.001*$
(PP) RKI SEES INFECTED'S CONTACTS TO INFORM ME	1.20	[0.98, 1.41]	$< 0.001*$
(PP) HACKERS KNOW INFECTION STATUS	−1.48	[−1.69, −1.27]	$< 0.001*$
(PP) RKI SEES DISTANCE VIOLATION	−0.87	[−1.09, −0.65]	$< 0.001*$
(PP) USES MY LOCATION TO PROTECT OTHERS	−0.10	[−0.33, 0.12]	0.359
(PP) UNNECESSARY QUARANTINE DUE TO FALSE POSITIVE WARNING	−1.34	[−1.55, −1.13]	$< 0.001*$
(PP) UNNECESSARY TESTING DUE TO FALSE POSITIVE WARNING	−0.87	[−1.09, −0.66]	$< 0.001*$
(PP) WARNING RESULTS IN QUARANTINE ENFORCEMENT	0.04	[−0.18, 0.27]	0.709
(PP) HAS DATABASE OF INFECTED AND CONTACTS	0.18	[−0.04, 0.41]	0.105
(PP) DATA PROTECTED BY NEW LAW	0.73	[0.52, 0.95]	$< 0.001*$
(PP) DATA PROTECTED BY GDPR	0.88	[0.67, 1.10]	$< 0.001*$
(PP) TECHNICAL PROTECTION OF DATA	1.00	[0.79, 1.22]	$< 0.001*$
(PP) TESTED BY BSI	0.90	[0.69, 1.12]	$< 0.001*$
(PP) TESTED BY IT EXPERTS	1.12	[0.91, 1.33]	$< 0.001*$
(PP) LOCATION NOT COLLECTED	1.10	[0.88, 1.31]	$< 0.001*$
(PP) CODE IS OPEN SOURCE	−0.12	[−0.34, 0.10]	0.277
(PP) ONLY LAW PREVENTS SURVEILLANCE	−0.53	[−0.75, −0.31]	$< 0.001*$
Neg. change | No change	−1.89	[−2.05, −1.74]	$< 0.001*$
No change | Pos. change	1.33	[1.31, 1.35]	$< 0.001*$

Table 4: Ordered logit regression model correlating different app properties against a combined "Usage Intention Change" scale ranging from 'Negative change" to "Positive change".

Factor	Description	Baseline
Required		
Trust in Government	5-point scale. Fully trust to fully distrust towards the government.	Neither
Optional		
Beliefs	3 multi-choice questions. Beliefs about the app in general, personal context, and related to others.	n/a
Worries	3 questions; 4-point scales. How worried are participants regarding future health, economy, and social life.	Not worried
Media Sources	Multi-choice question. From which media sources participants learned about the app.	n/a
Personal	6 questions; Yes, No & "Don't know". Health risks, previous infection, deaths, and other personal effects.	No
Demographics	7 questions. General demographic questions such as tech background, age, gender, and job.	various

Table 5: Factor categories appearing in the candidate regression models. Model candidates always included the required factors and covered all possible combinations of optional factors. Final models were selected based on lowest AIC. Categorical factors are individually compared to their listed baseline.

Abbreviation	Question	True?
General attributes		
(GEN) SHARES MOTION PROFILE	The app shares a profile of my movement.	✗ [36]
(GEN) SHARES TEMPORARY IDS	The app shares temporary IDs and timestamps.	✓ [38]
(GEN) SHARE PHONE CONTACTS WITH GOVERNMENT	The app shares the names and phone numbers of my contacts with the government.	✗ [18]
(GEN) GOVERNMENT CAN TRACK ME	The app enables the government to see my current location.	✗ [36]
(GEN) THREATS PRIVACY	The app undermines my privacy.	-
(GEN) DETECTS NEARBY USERS	The app determines when other smartphones are nearby that are also using the app.	✓ [38]
(GEN) HELPS WITH TESTING DECISION	The app facilitates decision-making on who should be tested for COVID-19.	✓ [19]
(GEN) SHARES FITNESS DATA	The app shares fitness data.	✗ [9]
(GEN) SHOWS NEGATIVE INFECTION STATE	The app can be used to demonstrate to others that I am not currently COVID-19 positive.	✗
(GEN) FIGHTS DISEASE SPREAD	The app can help fight the spread of the COVID-19 virus.	✓ [38]
(GEN) CAN DONATE HEALTH DATA	Through the app I can donate health data to the RKI for research purposes.	✗ [9]
(GEN) RESTRICTS BASIC RIGHTS	The app infringes my fundamental rights.	-
(GEN) GOVERNMENT SEES DISTANCE VIOLATION	The app enables the government to see if people are not keeping a safe distance from others.	✗ [14]
(GEN) USES LOCATION SERVICES	The app uses location services (like GPS).	- [3, 14]
(GEN) MANDATORY USAGE	Usage of the app will be mandatory.	✗ [38]
(GEN) USES BLUETOOTH	The app uses Bluetooth.	✓ [38]
(GEN) NONE	None of the above applies	✗
Attributes if others are infected		
(OTH) INFORMS IF INFECTED NEARBY	The app notifies me when an infected person is located nearby.	✗ [14]
(OTH) INFORMS IF CONTACTED INFECTED	The app notifies me if I have had contact with an individual who later tested positive for COVID-19.	✓ [38]
(OTH) GOVERNMENT SEES QUARANTINE VIOLATION	The app enables the government to see if someone is not complying with quarantine orders.	✗ [14]
(OTH) NONE	None of the above applies	✗
Attributes if I myself am infected		
(SLF) DATA TRANSMISSION ONLY AFTER CONFIRMATION	A physician or the public health authority has to confirm my positive COVID-19 test result before the app sends data to the RKI.	✓ [14]
(SLF) INFORMS MY CONTACTS	The app informs other app users who have been close to me that they may have contracted the virus.	✓ [38]
(SLF) GOVERNMENT SEES QUARANTINE VIOLATION	The app enables the government to see if I am not complying with quarantine orders.	✗ [14]
(SLF) SHARES DATA CONTINUOUSLY	The app sends data continuously to the RKI.	✗ [14]
(SLF) NONE	None of the above applies	✗

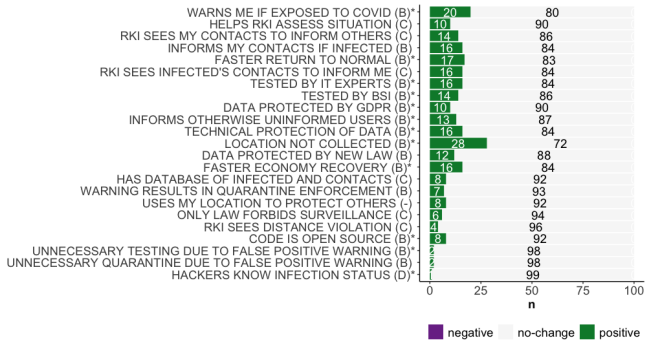
Table 6: Overview of all statements the participants were presented with and for which they had to decide whether they apply to the to be released CWA. The last column indicates if the attribute is correct for the app.

Abbreviation (Potential Property)	Full statement	True?	Approach (Central/ Decentral/Both)
(PP) ONLY LAW PREVENTS SURVEILLANCE	The government would be prevented by law, but not by technical means, from misusing the data for surveillance purposes.	✗ [28]	C
(PP) TECHNICAL PROTECTION OF DATA	Technical measures would be implemented to ensure the data are protected.	✓ [28]	B
(PP) RKI SEES INFECTED'S CONTACTS TO INFORM ME	If somebody near me has tested positive for COVID-19, the app would enable the RKI to see that I have had contact with that individual in order to notify me accordingly.	✗ [28]	C
(PP) RKI SEES MY CONTACTS TO INFORM OTHERS	If I test positive for COVID-19, the app would allow the RKI to see who I had contact with in order to notify those individuals.	✗ [28]	C
(PP) RKI SEES DISTANCE VIOLATION	Using the app would enable the RKI to find out if I am not complying with minimum distancing to other individuals.	✗ [28]	C
(PP) HAS DATABASE OF INFECTED AND CONTACTS	The RKI would have a database with the contact data of infected individuals and the people they have had contact with.	✗ [28]	C
(PP) HELPS RKI ASSESS SITUATION	The app would support the RKI to better assess the COVID-19 situation.	✗	C
(PP) USES MY LOCATION TO PROTECT OTHERS	The app would use information about my location to more accurately monitor infection risk for others.	✗ [14]	-
(PP) LOCATION NOT COLLECTED	The app would not collect any data about my location.	✓ [14]	B
(PP) TESTED BY BSI	The German Federal Office for Information Security(BSI) would verify that the app fulfills data security and data protection requirements.	✓ [4]	B
(PP) FASTER RETURN TO NORMAL	Using the app would make possible a speedier return to normal public life.	✓ [55]	B
(PP) UNNECESSARY QUARANTINE DUE TO FALSE POSITIVE WARNING	There is a possibility that the app could incorrectly report infection risk, resulting in me having to quarantine unnecessarily.	✗ [13]	B
(PP) FASTER ECONOMY RECOVERY	Using the app would help restart the economy faster.	✓	B
(PP) WARNING RESULTS IN QUARANTINE ENFORCEMENT	If the app notifies me that I may have been infected, I would have be required by law to quarantine.	✗ [13]	B
(PP) WARNS ME IF EXPOSED TO COVID	The app would notify me if I have been in a situation putting me at risk of contracting COVID-19.	✓ [38]	B
(PP) UNNECESSARY TESTING DUE TO FALSE POSITIVE WARNING	There is a possibility that the app could incorrectly report infection risk, resulting in me having to get tested unnecessarily.	✓ [19]	B
(PP) TESTED BY IT EXPERTS	Independent security experts would verify that the app fulfills data security and data protection requirements.	✓ [39]	B
(PP) DATA PROTECTED BY GDPR	Protection of the data would be guaranteed pursuant to a data protection policy and the General Data Protection Regulation.	✓ [28]	B
(PP) INFORMS OTHERWISE UNINFORMED USERS	The app would inform people of infection risk who would not otherwise be contacted by the public health authority.	✓ [38]	B
(PP) HACKERS KNOW INFECTION STATUS	Any nearby hackers could find out if I have tested positive for COVID-19.	✓ [47]	D
(PP) DATA PROTECTED BY NEW LAW	Protection of the data would be guaranteed under a new law drafted especially for the app.	✗	B
(PP) INFORMS MY CONTACTS IF INFECTED	If I have tested positive for COVID-19, the app would automatically notify other users of the app who are at risk being exposed through contact with me.	✗ [28]	B
(PP) CODE IS OPEN SOURCE	The app would be open-source	✓ [15]	B

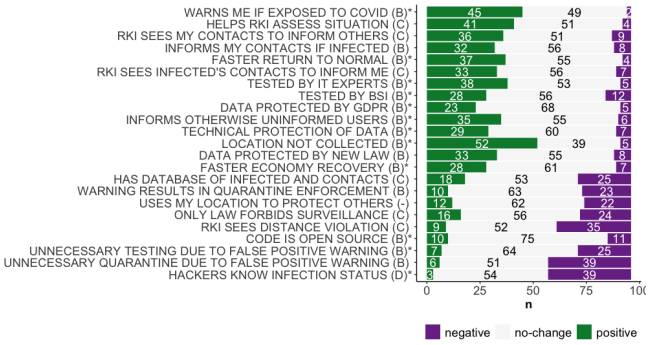
Table 7: The presented potential properties are either true for the centralized (C) or the decentralized (D) approach, or true for both (B) app designs. The properties that did not depend on the design approach is marked with "-".

Currently, how frequently do you have close personal contact with people not from your household?							
Once a week at most	39.5	A few times a week	37.8	A few times a day	10.2	Several times a day	10.1
Not disclosed	2.4						
How concerned or unconcerned are you about COVID-19 in regard to the following three areas?							
Health							
Unconcerned	16.0	A bit concerned	39.4	Concerned	25.7	Very concerned	18.3
Not disclosed	0.7						
The economy							
Unconcerned	7.3	A bit concerned	22.5	Concerned	34.5	Very concerned	35.2
Not disclosed	0.5						
Society							
Unconcerned	11.3	A bit concerned	23.9	Concerned	35.8	Very concerned	28.1
Not disclosed	0.9						
Do you fall within a COVID-19 high-risk group?							
Yes	31.1	No	58.1	Don't know	9.8	Not disclosed	1.9
Does someone close to you fall within a COVID-19 high-risk group?							
Yes	62.2	No	31.3	Don't know	5.8	Not disclosed	0.7
Have you or any person close to you fallen ill with Covid-19?							
Yes	7.5	No	86.8	Don't know	5.0	Not disclosed	0.7
Has anyone close to you died of Covid-19?							
Yes	3.0	No	94.9	Don't know	1.8	Not disclosed	0.4
How has the Covid-19 pandemic affected you financially?							
Positive impact	3.2	No impact	58.9	Negative impact	32.4	Critical impact	3.2
Not disclosed	2.3						
Has the Covid-19 pandemic resulted in you having to look after/care for someone at home?							
Yes	9.4	No	89.9	Not disclosed	0.7		
How has the crisis affected your work?							
Unaffected	52.7	Working from home	24.5	Short-time work	13.6	Became unemployed	5.0
Found employment	0.9	Not disclosed	3.4				

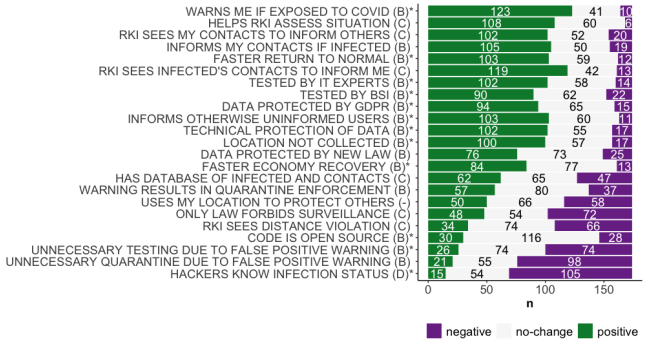
Table 8: Impact of the Covid-19 pandemic on participants. Numbers report the percentages in each question ($n = 744$)



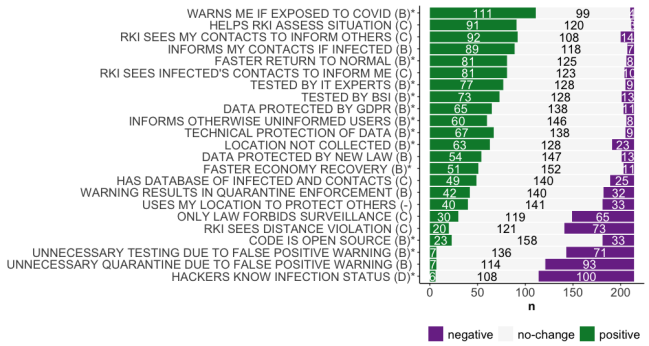
(a) Potential properties and the distribution of *Def-No* participants. $n = 100$



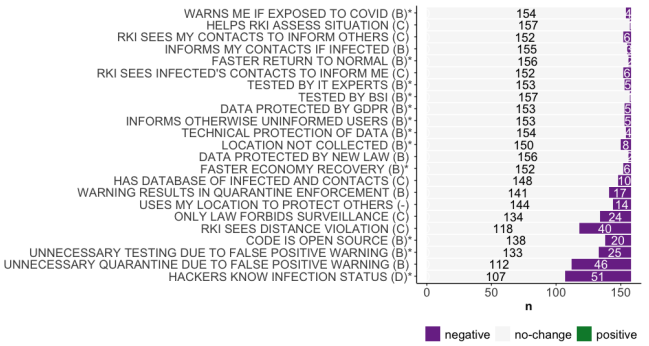
(b) Potential properties and the distribution of *Prob-No* participants. $n = 96$



(c) Potential properties and the distribution of *Undecided* participants. $n = 174$



(d) Potential properties and the distribution of *Prob-Yes* participants. $n = 214$



(e) Potential properties and the distribution of *Def-Yes* participants. $n = 158$

Figure 6: Participants perception of potential properties, spilt by their general usage intention. * indicate properties that apply to the real app. D = Dezantal, C = Central, B = Both

Understanding Users' Knowledge about the Privacy and Security of Browser Extensions

Ankit Kariryaa*
University of Copenhagen & University of Bremen
ak@di.ku.dk

Gian-Luca Savino*
University of Bremen
gsavino@uni-bremen.de

Carolin Stellmacher
University of Bremen
cstellma@uni-bremen.de

Johannes Schöning
University of Bremen & University of St. Gallen
schoening@uni-bremen.de

Abstract

Browser extensions enrich users' browsing experience, e.g., by blocking unwanted advertisements on websites. To perform these functions, users must grant certain permissions during the installation process. These permissions, however, give very limited information about the fact that they allow the extension to access user's personal data and browsing behaviour, posing security and privacy risks. To understand users' awareness of these privileges and the associated threats, we conducted an online survey with 353 participants, focusing on users' attitude, knowledge, and preference towards extensions' permission requests. We found that users report interest in seeking information, trust the developers but do little to protect their data. They have limited knowledge about the technical abilities of browser extensions and prefer permission statements that evoke a clear mental model. Based on our findings we derive recommendations for the improvement of browser extension permission dialogues through clear language, technical improvements and distinct responsibilities.

1 Introduction & Motivation

Web browsers are an important technology in modern daily life. We constantly use them to access online content for news, education, shopping or communication. As a result, browsers have a very large user base as well as a diverse scope of applications. To meet the requirements of such diverse use cases or enhance the browsing experience, browsers' functionalities can be extended through browser extensions.

Browser extensions, also known as browser add-ons, are small software programs that run inside a web browser. They are often developed by third-party companies or independent developers and are typically free of charge. Popular browsers have their web stores offering extensions in various categories ranging from productivity over accessibility to shopping (e.g. over 180k browser extensions are available for Google Chrome as of August 2019 [39]). The ten most popular browser extensions for Google Chrome alone have over 100 million combined downloads. The large number and variety of extensions enable users to customise their browser experience for their personal needs and preferences. Popular extensions block unwanted advertisements on websites or translate web text into the desired language. Others increase the accessibility of the browser through voice interaction [53] or automatically generate image descriptions on Twitter for people with vision impairment [35]. Recently, extensions were also proposed for detecting fake news through automatic fact checking [11], assisting users in the understanding of online privacy policies through spotting opt-out statements [7], or providing personalised password strength estimation [30].

To perform their intended purposes, extensions request permissions to access the content of visited websites and, often, other parts of the browser such as the browser's history. These special privileges enable extensions to read highly sensitive and personal data such as passwords or payment information, which can have serious implications for users' privacy and security. Especially with the nowadays ubiquitous online behaviour, knowledge about users' online browsing habits is highly valuable for revenue generation in the various domains (e.g., targeted advertisement). This illustrates a high commercial interest in users' browsing data which motivates malicious practices to gain access. Many leakage reports over the years [13, 49, 55] explored browser extensions and identified their role in online security and privacy issues.

To install an extension and benefit from its functionality users must grant all requested permissions and allow access to their browsing data. Users, therefore, have to make a trade-off between privacy concerns and convenience.

* denotes equal contribution.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

These trade-offs are not equally apparent to all users. As cybersecurity expert Schneier states in an interview:

> "In general, security experts aren't paranoid; we just have a better understanding of the trade-offs we're doing. Like everybody else, we regularly give up privacy for convenience. We just do it knowingly and consciously." [38]

It is, therefore, crucial to inform users appropriately about the data collection of browser extensions to enable them to make an informed decision. To do so, most browsers display a permission dialogue that users have to confirm to install browser extensions. However, explanations of these dialogues vary across browsers in regards to the user interface, used language and level of detail, as shown in Figure 1.

While it is easy and fast to install an extension at the click of a button and simultaneously grant the permission requests, it is unknown if the users are aware of the meaning and significance of these permissions and the associated risks. Recent research in the related topic of browsers' private modes has shown that differences in-browser explanations across browsers caused misconceptions about what private browsing mode does and how it protects users' privacy [54]. If such explanations can not convey the necessary knowledge, users are unable to make an informed decision and develop a false sense of security. To ensure their sovereignty over their personal data and to design better technology supporting them in privacy-related decisions, it is crucial to identify these gaps.

To better understand users' attitudes, knowledge, and preference towards browser extension permissions, we conducted an online survey with 353 participants. We investigated the effectiveness of modern browsers to communicate the meaning of permission requests. Our research particularly focuses on users' knowledge of the data that extensions can access and their understanding of the security and privacy risks coming with these privileges. We found that users have limited knowledge about the technical abilities of browser extensions. Their knowledge is mostly restricted to the beneficial features of the extensions they use and does not extend to other possible privacy and security risks. Their inability to apply this knowledge in a broader context shows their lack of technical understanding of the underlying permissions. Furthermore, users' perception of likelihood seems to be driven by the level of intrusion a scenario can potentially have on their privacy. We derive recommendations based upon our results and consider the perspective of users, developers and policymakers. These recommendations focus on improving the extension system, language of permissions and users' attitude. This paper contributes the first large scale survey on understanding users' attitudes, knowledge, and preferences about the privacy and security of browser extensions. Our study identifies a gap in users' perception about the current permission model and calls for long-overdue security improvements inspired by similar domains.

2 Related Work

As relevant prior work, we firstly summarise the background of the current browser extension system. We then present related research in the fields of human-computer interaction (HCI) and usable security about understanding users' knowledge, attitude and behaviour.

2.1 Browser Extensions & Browser Extension Security

With the exception of Safari[1], most modern browsers use the extension system that was first implemented by Google Chrome in 2009. The Chrome extension system stems from the design proposed by Barth et al. [8]. Their design was based upon the assumption that extension developers have good intentions but are, usually, not security experts. They argued that well-intentioned extension developers often write buggy code that can be exploited by malicious website operators to gain control over the extension. These exploits posed significant threats for two main reasons: 1) Under the former Firefox extension system, which was popular at that time, extensions often used unnecessarily powerful APIs and 2) they could have access to full user privileges at par with browsers or other native applications. To overcome these challenges, Barth et al. proposed a new browser extension system that improved the security of extensions by using principles of least privilege, privilege separation, and isolation. Their design separated the extension into three components, namely a content script, an extension core, and a native binary. Only the least privileged part of the extension (i.e. content scripts) was exposed to potentially malicious websites. In an evaluation of this security architecture, Carlini et al. [12] found it was mostly successful at preventing direct web attacks on extensions, but underlined its susceptibility to network attacks and website metadata attacks.

The Chrome extension system was designed to protect buggy-but-benign extensions, however, it provides no protection to users against intentionally malicious extensions. In recent years, a large number of browser extensions were found to be malicious, challenging the buggy-but-benign assumption. In an analysis between 2016 and 2018, Chen and Kapravelos identified over 3000 browser extensions from Chrome and Opera that were potentially leaking privacy-sensitive information [13]. The ten most popular Chrome browser extensions on that list, with confirmed malicious behaviour, affected over 60 million users. Another large-scale study investigated the 10,000 most popular browser extensions of Google Chrome and found that hundreds of extensions leaked sensitive information about users' browsing habits [49]. They found that while most extensions leaked information accidentally, e.g., when third-party content is injected into a website,

[1] Apple announced in WWDC 2020 that Safari will switch to the same extension API in the near future

others abused their access to user data on purpose. In July 2019, Jadali identified eight browser extensions with a total of 4 million downloads that collected browsing histories and exposed them in real time [28]. Similarly, a report from May 2020 identified 111 malicious extensions that were siphoning personal data such as passwords, credential tokens stored in cookies or parameters, screenshots, and tracking users browsing history [25]. Jointly, these 111 extensions had more than 32 million downloads. Another malevolent practice performed through browser extensions is malvertising which includes altering web content and displaying malicious advertisements, leading users to download and install malware. A screening of 18,000 Chrome extensions in 2015 found that extensions practising malvertising had over half a million users [55]. These studies and reports on malicious extensions with millions of downloads strongly challenge the assumption that all extension developers have good intentions.

To protect users from malicious extensions, most web stores use an automated review process [3, 15, 17, 22]. In certain cases, especially when sensitive permissions are involved, a manual review may follow an automated one. However, data leaks and reports of malicious activities underline the limits of these approaches. In response to privacy-breaching browser extensions, various solutions were proposed to protect or inform the users. These solutions include privacy-focused extensions to notify the user if an installed extension was suspected of malicious practices [51] or generating visits to random websites to conceal users' true browsing behaviour [49]. Similarly, to protect users from malvertising, Xing et al. proposed a browser extension that automatically detects extensions that inject ads [55]. These privacy-focused extensions not only increase users' privacy but can also improve users' browsing experience [10]. The proposed solutions and the review process of the web stores can assist users in protecting their online privacy and security. Nonetheless, the decision about extensions' security cannot be left to trusted parties alone [24]. The bulk of potential risks and the responsibility to make an informed choice lies with the user. The current practice of browsers to inform users about their extensions is using dialogues to describe the requested permissions during the installation process. However, there is limited research on users' attitude and knowledge towards browser extensions and the effect that permission dialogues have on them.

2.2 Privacy Knowledge and Data Sharing Behaviours

Next, we discuss related works in associated domains about understanding users' attitude and knowledge towards permissions and data collection.

In the domain of mobile applications, researchers found that smartphone users are often unaware of the permission settings and data collection of apps running on their devices [2, 9, 34, 48]. This is partly because users display low attention and comprehension when it comes to reading permission dialogues. In a study, Felt et al. found that 17% of participants paid attention to permissions during installation in a laboratory setting, and only 3% could correctly answer permission comprehension questions in an online survey [19]. When confronted with real app behaviours users felt their personal space had been violated [48]. This insight has led to studies trying to improve users' understanding of certain permissions and the data they give away. Almuhimedi et al. used a custom permission manager to make users aware of the data that applications were accessing and were able to make users reassess and restrict the permissions they were giving to applications [2]. Similar studies have also been conducted in other domains that deal with highly sensitive health data, such as wearable and fitness trackers [23, 40]. Research finds that with wearable technology it is less about the knowledge that data is collected, but about the value of this data [1] and the severity of the consequences of it being collected [47]. Schneegass et al. showed that non-expert users lack an understanding of the relationship between access to sensor data and access to information derived from this sensor data [47]. Furthermore, Aktypi et al. found that users highly underestimate the value of personal fitness data for third parties [1].

These studies across domains have been beneficial in developing systems that support users in making informed and sensible decisions with regards to access permissions. Even though browser extensions have existed for longer than mobile apps and fitness trackers, there is limited research in understanding users' attitude, knowledge or preference towards extension permissions, or making the extension permissions more understandable and usable. To fill this gap in the literature, in this paper, we study the privacy and security attitude of users towards browser extensions. We assess their knowledge about permissions and finally gather their preferences towards existing permission statements.

3 Method

We conducted an online survey to learn about (1) users' attitudes towards privacy and security topics related to browser extensions, (2) their general knowledge about browser extensions, (3) the influence of browser extension permission dialogues on their understanding, and (4) their preference for specific browser extension permission dialogues.

3.1 Browsers and Browser Extensions

In this paper, we study the browser dialogues of the most common browsers and browser extensions. As per Statista, the six largest desktop browsers by market share are Chrome (69.42%), Safari (8.74%), Firefox (8.48%), Edge (3.45%), Internet Explorer (2.88%) and Opera (2.39%) [50]. We excluded Internet Explorer in our study because Microsoft ended development for the browser in 2016 and replaced it with Edge.

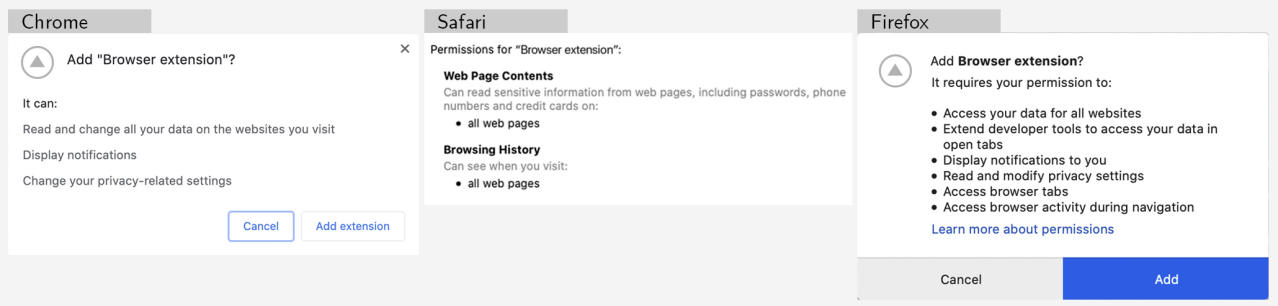


Figure 1: Permission dialogues of our sample extension for Chrome, Safari and Firefox. The sample extension used the super-set of permissions of the selected extensions.

Next, we surveyed the 10 most used extensions for each browser based upon download count if available, otherwise the number of ratings. Table 2 in the appendix (section B) shows these browser extensions for the five browsers. We, then, extracted the extensions with the highest appearance rate (number of browsers they appeared in) and ranked them by downloads across all browsers. Based on our criteria, we identified the following five extensions for our study: (1) Adblock Plus, (2) uBlock Origin, (3) Grammarly, (4) Adblock, and (5) Honey. Since Opera and Edge are both Chromium-based and their permissions are, therefore, almost identical, we decided to eliminate them from our study and focus on the differences between Chrome, Firefox and Safari (having around 86% market share). The permissions requested by the selected browser extensions are shown in Table 1. Next, we implemented a representative extension that used the super-set of permission requested by these five extensions in Chrome, Firefox, and Safari. Our extension used an ambiguous name and logo (see Figure 1). We locally installed the dummy extension on all the browsers and captured the actual permission dialogues.

3.2 Scenarios

To evaluate respondents' knowledge and beliefs about the technical abilities of browser extensions, we created ten scenarios. The scenarios were developed in iterative discussions involving three researchers with backgrounds in usable security and interface design (see section 4.3.2, figure 5 for a complete list of scenarios). Scenarios 1-3, 5, 8 were derived from the existing literature on malicious activities of extensions [13, 25, 28, 32, 37, 49, 55]. Scenarios 4, 6, 7, 9, 10 were added to ensure a broader possibility spectrum. The most common permission required by the browser extensions is to "Access all data on all websites". We found that among the 50 most downloaded browser extensions on the Firefox web store[2], 47 extensions request this permission. Given the ubiquitous need for this permission, five out of ten scenarios were based upon the functionality provided by it. Other scenarios considered access to the device's camera and microphone and the ability to control other extensions. Furthermore, three scenarios were based upon functionality outside the scope of a browser, namely, change the default password of the computer, restart the computer, and install an application on the computer.

We framed the scenarios as neutral statements without any harm being explicitly mentioned in them. We postulate that the neutral statements have a higher ecological validity as they can be considered direct derivatives of the statements of the browser permission dialogues. For example, scenario S_5 "The browser extension reads the user's usernames and passwords and stores them on an external server" is a specific case of the Chrome permission "Read and change all your data on websites you visit". Thus, the scenarios could test the case-specific knowledge of the various permission statements.

The permissions specified under the extension API allow the browser extensions to, among other things, access the web content, and access browsing history. In general, the permissions available under the extension API model are limited to the browsers. However, some extensions work in tandem with desktop applications such as Zotero[3] and Grammarly[4]. This model allows the extensions to leverage the privileges of their tandem applications and perform functions outside the scope of the extension API. Thus, in a broad sense, browser extensions can control any aspect of a computer, even though many functionalities are outside the scope of the extension API. In an absolute sense, all of the scenarios are technically possible but some require additional intervention by the user.

3.3 Survey Structure

Our survey comprised 35 unique questions, including attention checks. However, since it included randomisation and branching logic, the average participant was shown around 28 questions. The survey consisted of "yes/no/don't know",

[2] https://addons.mozilla.org/en-US/firefox/extensions/

[3] https://www.zotero.org/

[4] https://www.grammarly.com/

Extension	Downloads Chrome	Ratings Safari	Downloads Firefox										
				Chrome:	Read and change all your data on the websites you visit	Display notifications	Change your privacy-related settings						
				Safari:	Web Page Content: Can read sensitive information from web pages, including passwords, phone numbers and credit cards on: - all web pages			Browsing History: Can see when you visit: - all web pages					
				Firefox:	Access your data for all web sites	Display notifications to you	Read and modify privacy settings		Access IP address and hostname information	Store unlimited amount of client-side data	Access browser tabs	Access browser activity during navigation	Extend developer tools to access your data in open tabs
Adblock Plus	+10.0M	108	6.8M		C, S, F	C, F		S		F	F	F	F
uBlock Origin	+10.0M	*not available*	3.8M		C, F		C, F		F	F	F	F	
Grammarly	+10.0M	613	1.1M		C, S, F	C, F		S			F		
Adblock	+10.0M	1K	1.0M		C, S, F	C, F		S		F	F	F	F
Honey	+10.0M	4.6K	958K		C, S, F			S					

Table 1: Permissions requested by the selected browser extensions in Chrome (C), Firefox (F) and Safari (S). Download and rating count were retrieved in August 2020 from the respective browser extension stores.

multiple-choice, five-point Likert scale questions and one open-ended question. The complete survey can be found in the appendix (section A). Our survey methodology is adopted from similar studies in the HCI and usable security community that focused on understanding users' knowledge, attitude and behaviour for various digital platforms [23, 27, 44]. The survey consisted of the following sections:

Demographics: Participants' age and education as well as whether they have a professional background in any computer science-related field.

Confidence and attitudes regarding the information on browser extensions: The specific browsers and browser extension participants use. Their confidence about knowing what kind of data browser extensions collect and if developers made sure their data is safe. Their own precautions and attitudes towards privacy policies and terms and conditions.

Knowledge of the capabilities of browser extensions: The plausibility of the ten scenarios and the likelihood of them being used maliciously. Participants had to judge whether the scenarios were technically possible by answering "yes", "no", or "I don't know" and how likely they would deem the scenarios to be used maliciously on a five-point Likert scale from "very unlikely" to "very likely". In a separate question placed before the scenarios, we also asked the participants if an installed ad-blocker can read passwords on various websites.

Comprehension of extension permission dialogues: Comprehension and understanding of existing browser dialogues for Chrome, Firefox, and Safari. Participants were randomly presented with one of the browser extensions permission dialogues. They had to judge the same ten scenarios again on their plausibility and likelihood of being used maliciously, taking the permission dialogue into account. We were interested in whether the dialogues would convey the information to correctly assess the possibility of the scenarios if participants had previously failed to do so.

Preference of permission statements: The three browsers formulate their permission statement differently. Firefox uses all-inclusive words such as *"access"*, Chrome uses distinct keywords such as *"read and change"*, and Safari provides specific examples such as *"read sensitive information on web pages including passwords ..."*. For each of the browsers, we studied information conveyed and participants' preference for four commonly requested permissions. To do so, we created a comprehensive description including an explanation and examples for the four permissions. Our comprehensive description was based upon the reference text provided by the different browsers such as Firefox [20] and Chrome developer documentation [14]. To remove any bias towards a single source, we included the important keywords used by all browsers in our comprehensive description. For example, the following description represents the permission about access to information on all pages:

> "The browser extension can access, meaning read and change, all information including sensitive information such as passwords, phone numbers, credit card numbers, text and images on all websites such as those for online banking, email service, online shopping, and social media."

Participants were asked to rate the similarity and preference of browsers' original permission statements compared to the

comprehensive descriptions. The complete list of comprehensive descriptions that we developed for the study can be found in the survey (appendix section A, Q 6.1-4, page 17-18).

3.4 Participants

The recruitment of participants was done through the online platform Prolific [41]. The survey was hosted on Qualtrics [43]. 408 participants completed the survey, and on average it took them 11 minutes to finish it. The study participants had an average Prolific approval rating of 98.9% and they resided in more than 20 countries (mostly EU). Participants were paid £1.2 at a rate of £6.5 per hour. We excluded 12 participants due to failed attention checks and ended up with a total of 396 valid survey responses. Since we were interested in users of browser extensions, we excluded 43 participants from the main analysis who do not use browser extensions. We will, however, discuss their answers separately in our insights. Consequently, we analysed a data set of 353 responses of participants using browser extensions.

Of the 353 respondents, 219 (62%) identify as male, 134 (38%) as female. Their age ranged from 18 to 63 with a mean age of 28 ($SD = 9.6$). Regarding education, six (2%) had no formal education, 130 (37%) had a high school diploma or equivalent, 209 (59%) had a university degree and eight (2%) a doctoral degree.

3.5 Limitations and Ethical Considerations

While our study is based upon a relatively large and diverse sample, it may not be representative of the entire population. Our sample is relatively young, well educated and has a high proportion of people with a background in computer science. We also recorded 12 invalid responses from participants who left the survey early as well as 10 participants who did not finish the survey in the maximum allocated time of 49 minutes. This might have been due to a stereotype bias caused by the leading demographic questions, however, given the relatively small number of invalid responses the bias is unlikely to be pronounced. As it is sometimes the case with surveys in the domain of usable privacy and security, we would like to underline that some of our findings may have been impacted by social desirability and response bias where participants tend to present an inflated view of their privacy concerns, believing this is how the researchers want them to respond. Besides these, in our survey, most questions were based upon a rating scale and we had limited free text questions. For example, to study the appropriateness of extension browser permission statements, we asked the participants to evaluate their similarity and preference as compared to their comprehensive descriptions on a three-point scale (see section 4.3.5 and 5.2.4 for more information). While this approach allows us to determine the appropriateness of browser permission statements, it is not suitable to determine specific shortcomings in a given statement. A future work looking to elicit these specific shortcomings may find other approaches such as surveys focused on open response questions followed by qualitative coding more useful. It should also be noted that our study is based upon a frequently used, but limited, set of permissions, which only covers a part of the many permissions that are available to browser extensions.

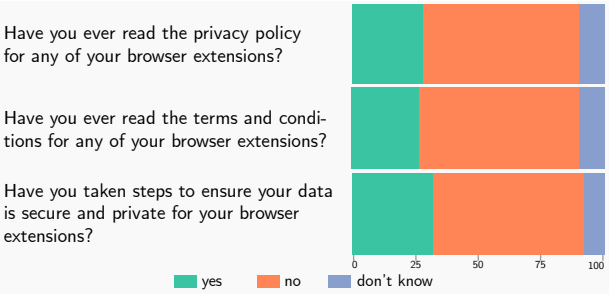


Figure 2: Participants' response to having read the privacy policy or terms and conditions of their installed extensions.

The survey was conducted within the ethical research guidelines of our university and did not require separate approval from the ethics board. Besides the Prolific IDs, which were necessary for compensating the participants, we did not collect any personally identifiable information in the survey.

4 Results

The following results were extracted from the survey and present users' usage of browsers and browser extensions, as well as our three main focal points on users' attitude, knowledge, and preference. Since our survey is exploratory, we primarily used descriptive statistics supported by graphic representations and complemented with significance testing where applicable.

4.1 Browsers and Browser Extensions

Most participants report Chrome (66%) as their default browser. This is followed by Firefox (18%), Opera (6%), Brave (4%), Edge (3%), and Safari (3%). Vivaldi, Yandex and Opera GX were also mentioned by one participant each.

85% of the participants use ad-blockers (e.g. Ad-block Plus, uBlock Origin), 30% use language tools (e.g. Oxford Dictionary, Grammarly), 29% use video or music downloaders (e.g. YouTube Downloader, Video DownloadHelper), 26% use password managers (e.g. LastPass, 1Password), 25% use shopping assistants (e.g. Honey, Piggy), and 19% use productivity tools (e.g. Todoist, Evernote).

Of the 43 respondents who do not use browser extensions, 44% didn't know they existed, 42% said that they do not need them, 7% find them too difficult to install, and 5% do not

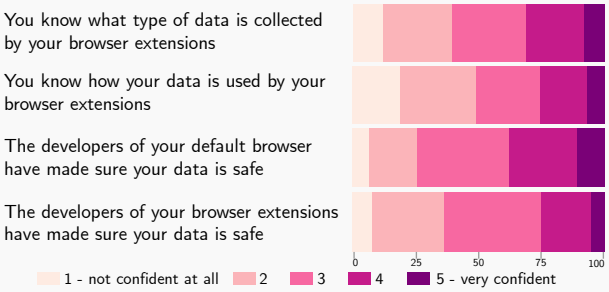


Figure 3: Participants' confidence in their own knowledge about the data collection, data usage and developers of their installed browser extensions.

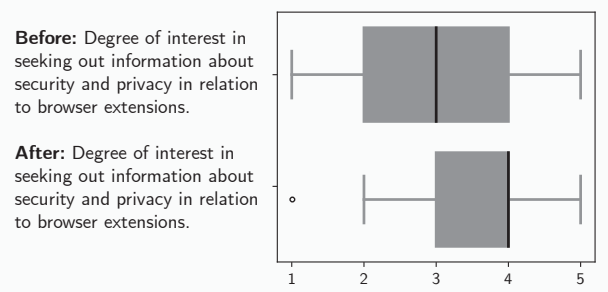


Figure 4: Interest in seeking information in the beginning and end of the survey.

use them because of concerns about data privacy. 2% were uncertain if they were using browser extensions.

4.2 Attitude

This section presents users' attitudes towards reading terms and conditions, their confidence in developers of web browsers and browser extensions, the impact of the permission dialogue and their interest in seeking information on security and privacy concerning browser extensions.

4.2.1 Terms and Conditions

More than 60% of the participants in our survey reported that they have not read the privacy policy or the terms and conditions of their installed browser extensions. Furthermore, 59% reported that they did not take any steps to ensure their data is safe with the browser extensions. Figure 2 shows the response of the participants.

4.2.2 Confidence in Developers

Figure 3 shows the participants' confidence in developers of their default browser and installed browser extensions, that they have ensured user data is not being tampered with or shared without explicit consent. Participants showed slightly higher confidence in the developers of their default browser ($median = 3.0, mean = 3.17, SD = 1.07$) compared to the developers of their browser extensions ($median = 3.0, mean = 2.87, SD = 1.00$). A Wilcoxon signed-rank test showed that the differences were statistically significant ($p < 0.001$). Only a small number of participants had either very high or no confidence in the developers of both browsers and browser extensions, with three out of five being the most frequent choice.

4.2.3 Awareness of Permission Dialogues

To study users' awareness of permission dialogues, we showed participants the Chrome dialogue as a representative of browser extension dialogues. 123 (34.8%) of the participants reported that they had seen the provided example or a similar permission dialogue before. 28.3% reported that they had not seen a permission dialogue, and the rest could not remember. Out of the 123 who had seen a permission dialogue, 68% reported that it influenced their decision about installing the browser extension.

4.2.4 Interest in Seeking Information

We asked the participants about their interest in seeking out information on security and privacy concerning browser extensions in the beginning as well as at the end of the survey. At the end of the survey, participants were more interested in seeking out information and the median interest increased from three to four. The results are shown in figure 4. A Wilcoxon signed-rank test showed a statistically significant difference between the interest in the beginning and the end of the survey ($p < 0.001$). Furthermore, most participants are not comfortable in having browsing history or personal data being collected and stored by a browser extension. On a scale from 1 - Not at all comfortable to 5 - Extremely comfortable, they gave a median score of two.

4.3 Knowledge

In this section, we report our findings in regards to participants' knowledge about browser extensions, the practices of data collection and the impact permission dialogues have on users' knowledge.

4.3.1 Data Collection And Use

We asked participants to rate their confidence in their knowledge of what data is collected, and how the collected data is used by browser extensions. Participants rated their confidence on a five-point unipolar Likert scale from 1 - Not at all confident to 5 - very confident. Figure 3 shows a mostly uniform distribution of the responses. However, participants were less confident in how the data is used than what type of data is collected.

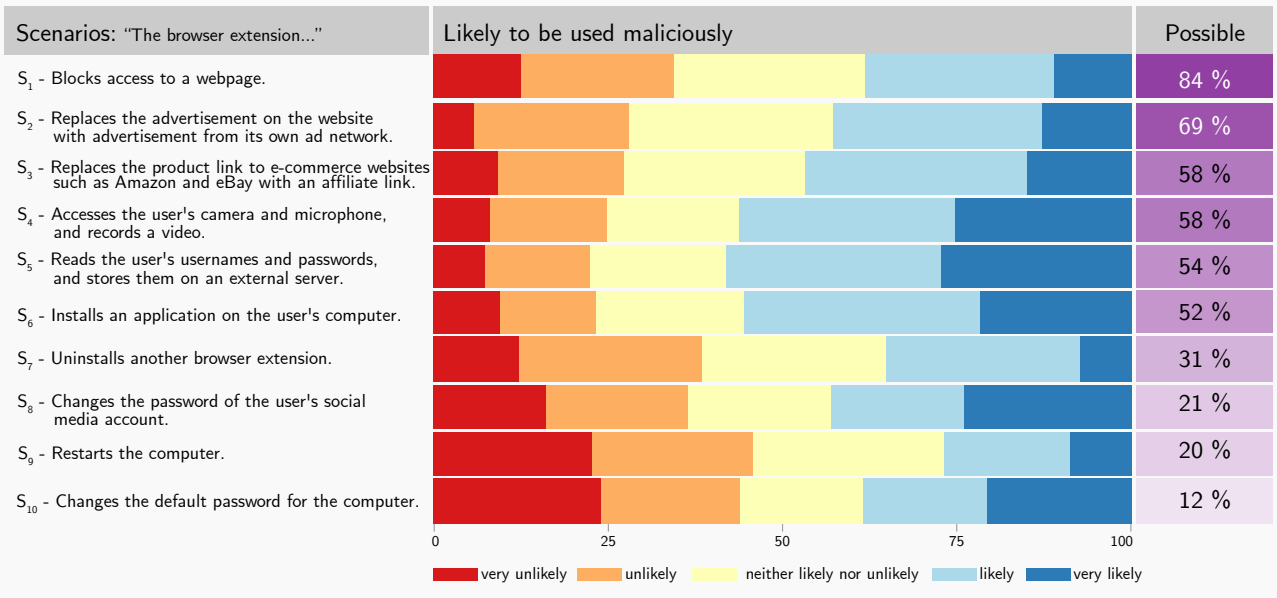


Figure 5: Participants' perception of the plausibility of scenarios and likelihood of them being used maliciously.

4.3.2 Users' Knowledge of Browser Extensions

For each of the ten scenarios (see figure 5), we asked participants to select "yes", "no", "don't know" to indicate whether they thought a scenario was technically possible to occur (which all of them were as explained in section 3.2). We found that most users thought S_1: "Blocks access to a web page" (84%) and S_2: "Replaces advertisements" (69%) was possible. Roughly half of all participants thought S_3: "Replaces product links" (58%), S_4: "Accesses the camera and microphone" (58%), S_5: "Reads the user's password" (54%), and S_6: "Installs an application on the user's computer" (52%) were possible to occur. Less than a third of the participants thought S_7, S_8, S_9, S_{10} were possible. Figure 6 shows the reported plausibility of selected scenarios across the conditions.

Furthermore, participants rated the likelihood of the scenarios being used maliciously. On a bipolar five-point Likert scale, the median response was "likely" for scenarios S_4, S_5, and S_6, and "neither likely nor unlikely" for the rest. The impact of the different permission dialogues on participants' perceptions is further illustrated in a graph in the appendix (figure 8).

4.3.3 Knowledge of Ad-Blockers

To the separate question on "Assuming that you have an ad-blocker installed as a browser extension, can it read passwords that you use on various websites?", 41 (9%) participants selected "yes", 142 (40%) selected "no" and rest of the participants (51%) did not know.

4.3.4 Effectiveness of The Browser Extension Dialogues

To test how effective the browser dialogues were in communicating the abilities of the browser extension, we scored participants' knowledge before and after they saw the dialogue. To calculate the score, one point was added for a correct assessment of a scenario to be possible, one point was subtracted for a wrong answer, and no point was added for answering "I don't know". For this comparison, we only took scenarios S_1, S_2, S_3, S_5, and S_8 into account which were explicitly permissible by the permissions (i.e. without the need of a tandem application). Thus, the maximum score was $+5$ and the minimum score was -5. Without seeing a dialogue, respondents had a median score of two ($mean = 1.69, SD = 2.48$). After seeing a dialogue the median score increased significantly ($p = 0.015$) to three ($mean = 2.05, SD = 2.68$) across all browser dialogues. Regarding individual browsers, participants who saw the Firefox dialogue had a median score of three ($mean = 2.16, SD = 2.54$), those who saw the Chrome dialogue had a median score of three ($mean = 2.6, SD = 2.41$), and the ones who saw the Safari dialogue had a median score of two ($mean = 1.29, SD = 2.96$). A Kruskal-Wallis test found a significant difference between the browsers ($p < .001$). Post-hoc Wilcoxon rank-sum tests found that there is a statistically significant difference between Firefox and without dialogue ($p = .045, r = .09$), Chrome and without dialogue ($p < .001, r = .17$), Firefox and Safari ($p = .038, r = .14$), and Chrome and Safari ($p < .001, r = .22$). Effect sizes were calculated according to Robertson and Kaptein [45]. To summarise, Chrome and Firefox significantly improved the score as compared to the baseline (i.e. without-dialogue) and Safari with a small effect.

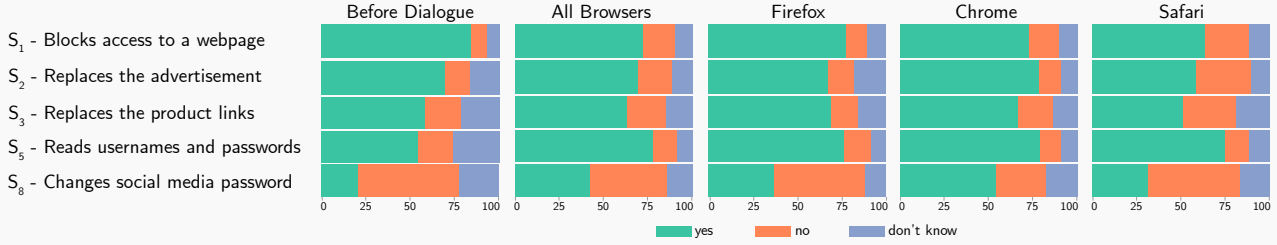


Figure 6: Participants' evaluation of the plausibility of selected scenarios. Each participant first evaluated the plausibility without seeing a permission dialogue, and then again after seeing a permission dialogue of either Firefox, Chrome, or Safari (randomly chosen). The permission dialogues for the three browsers are shown in figure 1.

4.3.5 Similarity and Preference of Existing Permission Dialogues

Participants rated all permission statements as similar to our comprehensive descriptions. Figure 7 shows that for different statements, the majority of participants rated them "Extremely similar", and less than 13% rated them to be "Not at all similar". However, participants did not prefer most of the original browser permission statements to be used instead of our description. For all statements except two, the majority of the participants rated them as "Not at all preferred", and less than 12% rated them to be "Extremely preferred". The only two statements for which participants reported slightly higher preference were "display notification" and "display notification to you" (see figure 7).

5 Findings

We draw the following main insights from the findings of our survey results.

5.1 Attitude

We found that the majority of users are interested in seeking out information about security and privacy in relation to browser extensions. They feel somewhat confident about what data is collected by their browser extensions and how the data is used. However, less than a third have ever read the terms and conditions or the privacy policy of their browser extensions or have taken any steps to ensure their private data is secure. Here our findings are in line with existing literature that the majority of the users do not read privacy policies [5, 23] and further highlight the low utility of terms and conditions and privacy policies in conveying to the user what information online services collect and how it is used [26, 29, 36].

5.1.1 Trust in Developers

With regards to the access and storage of users' data, the majority of the participants reported moderate to high trust in developers. They put slightly higher trust in developers of the browsers as compared to the trust in the developers of extensions. Here the results vary from our initial hypothesis that the trust in developers of browsers would be notably higher as compared to the trust in the extension developers since, in contrast to browser extensions, browsers are universally adapted applications and developed by selected organisations. Given the findings, we speculate that the trust in browsers is extended to the trust in the browser extensions since browser extensions are distributed through the browsers' webstore.

5.1.2 Users Seek More Information

Our results show that after having completed our survey, participants' interest in seeking more information about security and privacy in regards to browser extensions increased significantly. In the text field at the end of the survey P54 commented "I'm more aware of the risks now". P311 wrote: "It made me more aware of the vulnerabilities of all the extensions I use". For some participants, the survey even made them reiterate their past and future decisions (P85): "I will re-read all my extensions and read the terms every time I install a new one".

5.2 Knowledge

5.2.1 Users' General Knowledge about Technical Abilities of Browser Extensions is Limited

Participants gave higher possibility ratings to scenarios that are closely related to specific types of browser extensions. The two scenarios where participants were most sure of their plausibility are: S_1: "Blocks access to a web page" (84%) and S_2: "Replaces advertisements" (69%). Both scenarios are strongly connected to ad-blockers which the majority of participants use (85%). Scenario S_3: "Replaces the product link" is most likely associated with shopping assistants and scenario S_5: "Reads user names and passwords" with Password managers, which both a quarter of all participants use.

Regarding the likelihood, scenarios were rated more likely to be used maliciously when they included an invasion of privacy. S_4: "Accesses the camera and microphone", S_5: "Reads

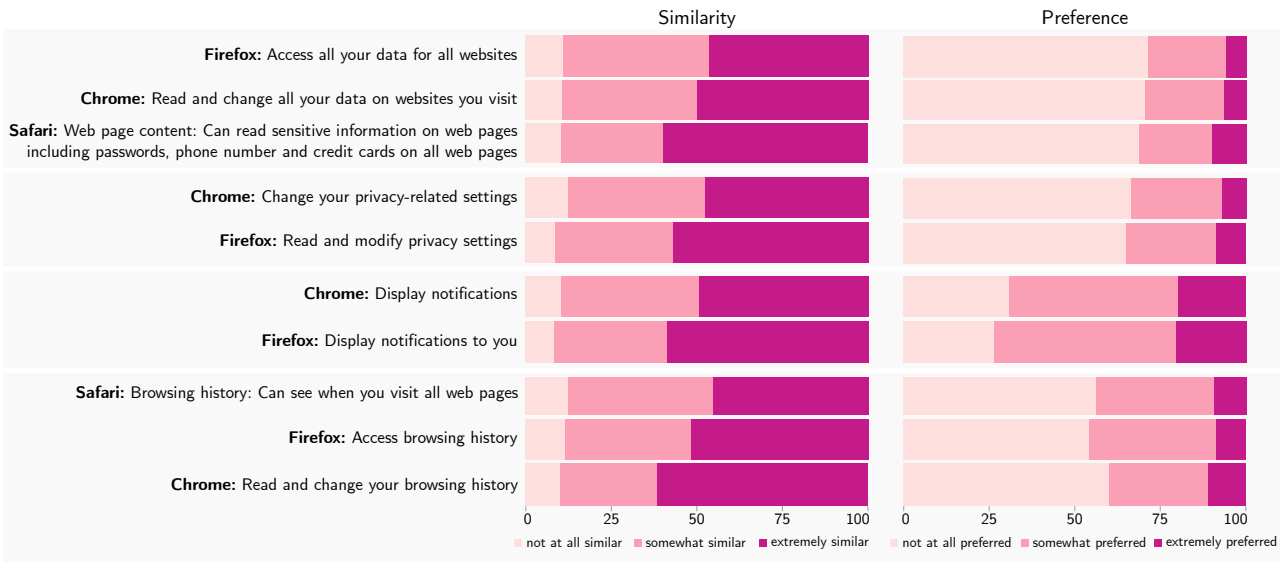


Figure 7: Participants' rating of similarity and preference of existing permission statements in comparison to our comprehensive descriptions. The majority of the participants rated existing permission statements to be similar but did not prefer them in place of our description.

the user's password", and S_6: "Installs an application on the user's computer", had the highest ratings in this regard.

Generally, we observe that users' understanding relies on individual experiences with specific extensions. While users seem to be aware that browser extensions such as password managers can read and store passwords, they do not consider that similar permissions are enabling ad-blockers to do the same.

5.2.2 Permission Dialogues Have Limited Effectiveness

We found significant differences between the browsers and observed that the framing of the permission statements matters just as much as having permission statements in the first place. The dialogues of Chrome and Firefox significantly improved participants' scores as compared to Safaris' dialogue and the baseline. Overall, the browser permission dialogues improved participants' scores significantly. While the median score improved from two to three across all browsers, 30% of all participants still scored 0 and lower. Even with the permission dialogues, participants were still not entirely informed about the technical implications of all the permissions.

5.2.3 Specific Statements Restrict Peoples' Ability to See Implications

Both Chrome and Firefox's permission dialogues improved user scores as compared to the baseline and Safari. Scenario S_5 "Reads usernames and passwords" was the only scenario where the Safari condition shows similar plausibility scores to Chrome and Firefox. This was also the only scenario explicitly mentioned in Safari's permission dialogue (see figure 1). The same permission, however, also enables the other four scenarios, which users, seeing the Safari permission dialogue, did not consider to be possible. Contrary to our initial hypothesis, that permission statements with examples would improve users' overall understanding, we find that Safari's dialogue statements are too specific and limit users' ability to see its implications for other scenarios.

5.2.4 Users do Not Prefer Existing Permission Statements

In the survey questions Q6.1-4, we had asked participants about the similarity in the information conveyed by four existing permission statements, as well as their preference for these browser permission statements compared to the comprehensive descriptions. When analysing the response of the participants we consider the four cases by dividing the score into low and high for similarity and preference. Firstly we consider the case when most participants give high similarity and high preference scores to the browser permission statement compared to its comprehensive description. We argue that this implies that the two statements have the same meaning and that the comprehensive description is a natural elaboration of the browser permission statement. In this case, the browser permission statement is better (and thus more preferred) since the users do not gain new information from the comprehensive description. We observe this case in response to question Q6.3 for the "send notifications" statement.

In the second case, most participants give a high similarity but low preference score to the browser permission statement.

This would imply that the comprehensive description can be compacted to the browser permission statement but with loss of information. In this case, the comprehensive description is better since the user gains new and relevant information from it (and thus the browser permission statement is less preferred). We observe this case in response to question Q6.1, Q6.2 and Q6.4 for "access all data for all websites", "change privacy related settings" and "access browsing history" related statements for the three browsers. Our results suggest that existing permission dialogues for these three permissions are too limited to be regarded as a suitable representative of their underlying meaning. The other two cases with a low similarity score would imply that the statements are disjoint. These cases are not observed in the responses.

5.3 Summary

Our findings show that users have a conflicting attitude towards privacy and security topics when it comes to browser extensions. Although users indicate interest in the topic, the majority of them have not read privacy policies, terms and conditions or taken steps to ensure the safety of their personal data. They trust developers to securely handle their data but often lack the knowledge to see the potential threats. Users' knowledge in regards to browser extensions is highly connected to individual experiences. While most users know extensions can read passwords, probably because of their experience with password managers, they don't consider that similar permissions enable ad-blockers to do the same. Permission dialogues help users, but unfortunately, their effectiveness is limited in building this understanding. We find that they alone are not sufficient to impart the knowledge needed for making informed decisions. Finally, we see preference as an important lever to make information accessible to users. The improved knowledge can change peoples' attitude towards topics such as security and privacy of browser extensions.

Overall we conclude that people's attitudes can be positively affected through knowledge as our evaluation about users' interest in the topic, before and after the survey, suggests (see figure 4). Participants were not only more interested but also more concerned about the topic. P288 even commented: "I am now scared of browser extensions", which highlights the scale of the problem and the large gap that is there to close in users' understanding of browser extensions.

6 Discussion

Browser extensions use extensive permissions, such as one single permission to access a whole website. This coarse model allows a range of extensions to access all kinds of user data. For example, under this model, both ad-blockers and dictionary extensions can technically access the same sensitive user information such as passwords, tokens, page URL, and payment information. To inform about these possibilities, browsers present the requested permissions during the installation process to assess the technical abilities of the extensions before (re-)confirming users' decision of installing an extension. These permissions aim to assist users to fill the knowledge gap between the technical abilities and functionalities of an extension.

In this paper, we explored the effectiveness of browser permissions in informing users about the plausibility of the technical abilities of the extensions. We find that the current model leaves a gap between the conveyed information and the user's understanding. In many cases, users have a misconception about the technical abilities of the extensions and the majority does not think that these abilities are likely to be used maliciously. More importantly, the permission statements have limited success in conveying the technical abilities of the extensions or changing users' perception of the likelihood. The problems are further exacerbated because the majority of the users do not read the privacy policy or the terms and conditions, and they have moderate to high trust in the browsers and browser extensions.

The results of our study are in line with trends in the related domain of mobile applications. We find that the majority of the users have a considerable understanding of the scenario related to popular features, such as the ability of an ad-blocker to replace advertisement or block access. However, they lack understanding about other scenarios feasible under the same permissions. Similarly, studies in the fields of mobile applications and wearables have shown that users have a limited understanding of the permissions and data collection [1, 2, 6, 19, 23]. In our study, only 34% of participants recall seeing the permission dialogue. This is similar to the results of Felt et al. on users' behaviour towards mobile app permission dialogues, where they found that the majority of the people just skip over or accept them without reading [19].

Even though in many ways, browser extensions are more powerful than and equally popular as mobile applications, limited research has looked into users' understanding of their abilities and users' attitude towards them, while numerous studies have been conducted for mobile applications [4, 18, 19, 33, 56]. With this study, we take a step towards filling this gap in the usable privacy and security literature for browser extensions.

Over the years, various measures for data security have been proposed, such as the *right of informational self-determination* introduced by German Federal Constitutional Court, which states the users should be able to decide what parts of their personal data can be accessed by whom, Canada's Personal Information Protection and Electronic Documents Act (PIPEDA) [42] and EU's General Data Protection Regulation (GDPR) [52]. In this paper, we present evidence that, with respect to browser extensions, users cannot make use of these rights, because they do not know that data is made accessible, let alone what kind of data is given away.

Users seem unaware of the trade-off between privacy and convenience and they are unable to make an informed choice in relation to browser extensions.

6.1 Recommendations

Based on our findings, we provide recommendations for further developing the extension model to empower the users and to safeguard them from malicious extensions.

1. Users' perspective: Improve understanding

The overall goal of the permission dialogues should be to increase informational self-determination by users. We find that existing permission statements are not a suitable representative of their underlying meaning in case of three out of the four permissions that we studied. Considering these aspects, we recommend that the browser should not assume that users are aware of the meaning of various permissions, and instead encourage the user in gaining information about the extent of the permissions. Here we suggest the use of "Human in the loop" framework [16] while designing the permission dialogues to ensure that the users comprehend the meaning of the statements. As Cranor [16] points out, similarity to related symbols, complexity and vocabulary all impact comprehension. Using familiar and unambiguous statements can aid these shortcomings and offer help in building understanding for terms that do not yet have a stable mental model. Similarly, for conveying which information is collected and how it is processed, an approach based upon the "nutrition label for privacy" may be better suited [31].

Building on existing experiences, we also recommend *parity with similar systems:* Recently, iOS 14 and Android 10 introduced a new fine-grained permission system that allows for time-limited permissions, permission reminders and confirmation on first-use, among others. Browsers already have runtime notifications for some permissions such as camera and microphone. While this may not be possible for all extension permissions as they are required for the basic functioning of the extension, the browser may still benefit from displaying permissions on runtime (i.e. when users first open a webpage) instead of only during the installation as it would allow the user to see the permission in the context of the webpage. Browsers can further call attention to the extent of the permission by highlighting the parts of the web page and browser settings accessible under it.

2. Browser's perspective: Limit access To improve the existing permission system, browsers should assume that deliberately or otherwise the extensions are prone to be malicious. Following the principle of least privilege [46], we recommend that browser extensions should be provided access only to the relevant part of the DOM. The sensitive information should be redacted from the extensions that do not need it (*Redacted DOM*). Most browsers already identify sensitive fields (such as password or credit-card fields) and, thus, they can be encapsulated with separate permissions.

Furthermore, as proposed in Chrome Manifest V3[5] and Apple WWDC 2020[6] browsers should provide the possibility to limit the scope of the extensions to certain categories of websites (*Restricted website access*). This feature could be especially helpful in preventing malicious extensions from gaining access to sensitive information on corporate websites or financial web services.

3. Policy perspective: Convey responsibility We recommend that browsers should make users aware of their responsibilities as well as the responsibilities browsers take on themselves. Browsers should convey to the users that they are responsible to only allow access to the extension APIs specified in the permission statements, and the users' responsibility lies in making an informed choice after knowing the upper bounds from the permission statements and understanding the actual behaviour through terms and conditions. If the browser takes additional responsibility they should explicitly specify it. A similar technique is adopted by the Firefox Recommended Extensions program to promote the safest and highest quality extensions [21]. We hypothesise that making users aware of their responsibility can improve their attitude towards online security in the long term. Further studies are required to establish the effectiveness of our recommendation on clearer language, parity with similar systems, redacted DOM, restricted website access, and responsibility conveying.

7 Conclusion

To conclude, our survey results have provided insight into the attitude, understanding and preferences towards security and privacy practices of browser extension users. Users expressed confidence in their knowledge of what data is collected and trust developers to securely handle their data but they have limited understanding to assess the potential risks. Users' knowledge in regards to browser extensions seems to be connected to individual experiences. For example, while most users know extensions can read passwords, probably due to their experience with password managers, they don't consider that similar permissions enable ad-blockers to do the same. Overall, our findings lead us to believe that browser extension users require a greater awareness of the risks associated with browser extensions and future work should look into making extension permissions understandable and fine-grained.

Acknowledgments

We would like to thank the anonymous reviewers for their constructive feedback that greatly improved the paper, Jasmin Niess for her help in designing the study, and Nadine Wagener for the help with the figures. This research was supported by

[5] https://developer.chrome.com/docs/extensions/mv3/intro/
[6] https://developer.apple.com/wwdc20/

the Volkswagen Foundation through a Lichtenberg Professorship and by the Federal Ministry of Education and Research of Germany (BMBF) through the Wintermute project (award number 16KIS1127).

References

[1] Angeliki Aktypi, Jason R.C. Nurse, and Michael Goldsmith. Unwinding ariadne's identity thread: Privacy risks with fitness trackers and online social networks. In *Proceedings of the 2017 on Multimedia Privacy and Security*, MPS '17, page 1–11, New York, NY, USA, 2017. Association for Computing Machinery.

[2] Hazim Almuhimedi, Florian Schaub, Norman Sadeh, Idris Adjerid, Alessandro Acquisti, Joshua Gluck, Lorrie Faith Cranor, and Yuvraj Agarwal. Your location has been shared 5,398 times! a field study on mobile app privacy nudging. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 787–796, New York, NY, USA, 2015. Association for Computing Machinery.

[3] Opera Software AS. Publishing Guidelines. Retrieved June 11, 2020 from https://dev.opera.com/extensions/publishing-guidelines/.

[4] Kathy Wain Yee Au, Yi Fan Zhou, Zhen Huang, and David Lie. Pscout: analyzing the android permission specification. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 217–228, 2012.

[5] Brooke Auxier, Monica Anderson Lee Raine, Andrew Perrin, Madhu Kumar, and Erica Turner. Americans' attitudes and experiences with privacy policies and laws, 2019. Retrieved May 20, 2021, https://www.pewresearch.org/internet/2019/11/15/americans-attitudes-and-experiences-with-privacy-policies-and-laws/.

[6] Rebecca Balebako, Jaeyeon Jung, Wei Lu, Lorrie Faith Cranor, and Carolyn Nguyen. " little brothers watching you" raising awareness of data leaks on smartphones. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, pages 1–11, 2013.

[7] Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Cranor, Shomir Wilson, Florian Schaub, and Norman Sadeh. Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text. In *Proceedings of The Web Conference 2020*, WWW '20, pages 1943–1954, New York, NY, USA, 2020. Association for Computing Machinery.

[8] Adam Barth, Adrienne Porter Felt, Prateek Saxena, and Aaron Boodman. Protecting browsers from extension vulnerabilities. In *Network and Distributed System Security Symposium*, 2010.

[9] Matthias Böhmer, Brent Hecht, Johannes Schöning, Antonio Krüger, and Gernot Bauer. Falling asleep with angry birds, facebook and kindle: a large scale study on mobile application usage. In *Proceedings of the 13th international conference on Human computer interaction with mobile devices and services*, pages 47–56, 2011.

[10] Kevin Borgolte and Nick Feamster. Understanding the Performance Costs and Benefits of Privacy-focused Browser Extensions. In *The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020*, WWW '20, pages 2275–2286, New York, NY, USA, 2020. Association for Computing Machinery.

[11] Bjarte Botnevik, Eirik Sakariassen, and Vinay Setty. BRENDA: Browser Extension for Fake News Detection. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 2117–2120, New York, NY, USA, 2020. Association for Computing Machinery.

[12] Nicholas Carlini, Adrienne Porter Felt, and David Wagner. An evaluation of the google chrome extension security architecture. In *Presented as part of the 21st {USENIX} Security Symposium ({USENIX} Security 12)*, pages 97–111, 2012.

[13] Alexandros Chen, Quan and Kapravelos. Mystique: Uncovering Information Leakage from Browser Extensions. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, pages 1687–1700, New York, NY, USA, 2018. Association for Computing Machinery.

[14] Google Chrome. Declare Permissions. Retrieved June 20, 2020 from https://developer.chrome.com/extensions/declare_permissions.

[15] Google Chrome. Publish in the Chrome Web Store. Retrieved June 11, 2020 from https://developer.chrome.com/webstore/publish.

[16] Lorrie Faith Cranor. A framework for reasoning about the human in the loop. In *Proceedings of the 1st Conference on Usability, Psychology, and Security*, UPSEC'08, USA, 2008. USENIX Association.

[17] Microsoft Edge. Publish Your Extension. Retrieved June 11, 2020 from https://docs.microsoft.com/en-us/microsoft-edge/extensions-chromium/publish/publish-extension.

[18] Adrienne Porter Felt, Erika Chin, Steve Hanna, Dawn Song, and David Wagner. Android permissions demystified. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 627–638, 2011.

[19] Adrienne Porter Felt, Elizabeth Ha, Serge Egelman, Ariel Haney, Erika Chin, and David Wagner. Android permissions: User attention, comprehension, and behavior. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, New York, NY, USA, 2012. Association for Computing Machinery.

[20] Mozilla Firefox. Permission Request Messages for Firefox Extensions. Retrieved September 11, 2020, https://support.mozilla.org/en-US/kb/permission-request-messages-firefox-extensions.

[21] Mozilla Firefox. Recommended Extensions program. Retrieved June 11, 2020 from https://blog.mozilla.org/firefox/firefox-recommended-extensions/.

[22] Mozilla Firefox. Submitting an Add-on. Retrieved June 11, 2020 from https://extensionworkshop.com/documentation/publish/submitting-an-add-on/.

[23] Sandra Gabriele and Sonia Chiasson. Understanding fitness tracker users' security and privacy knowledge, attitudes and behaviours. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–12, New York, NY, USA, 2020. Association for Computing Machinery.

[24] Cristiano Giuffrida, Stefano Ortolani, and Bruno Crispo. Memoirs of a Browser: A Cross-Browser Detection Model for Privacy-Breaching Extensions. In *ASIACCS 2012 - 7th ACM Symposium on Information, Computer and Communications Security*, ASIACCS '12, pages 10–11, New York, NY, USA, 2012. Association for Computing Machinery.

[25] Gary Golomb. The Internet's New Arms Dealers: Malicious Domain Registrars. Retrieved June 11, 2020 from https://awakesecurity.com/blog/the-internets-new-arms-dealers-malicious-domain-registrars/.

[26] Cormac Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *Proceedings of the 2009 workshop on New security paradigms workshop*, pages 133–144, 2009.

[27] Chris Jay Hoofnagle, Jennifer King, Su Li, and Joseph Turow. How different are young adults from older adults when it comes to information privacy attitudes and policies? *Available at SSRN 1589864*, 2010.

[28] Sam Jadali. DataSpii: The Catastrophic Data Leak via Browser Extensions. Retrieved July 7, 2020 from https://securitywithsam.com/2019/07/dataspii-leak-via-browser-extensions/.

[29] Carlos Jensen and Colin Potts. Privacy policies as decision-making tools: an evaluation of online privacy notices. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 471–478, 2004.

[30] Ankit Kariryaa and Johannes Schöning. Moiprivacy: Design and evaluation of a personal password meter. In *19th International Conference on Mobile and Ubiquitous Multimedia*, MUM 2020, page 201–211, New York, NY, USA, 2020. Association for Computing Machinery.

[31] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W Reeder. A" nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, pages 1–12, 2009.

[32] Swati Khandelwal. 8 More Chrome Extensions Hijacked to Target 4.8 Million Users. Retrieved May 25, 2021 from https://thehackernews.com/2017/08/chrome-extension-hacking.html.

[33] Jialiu Lin, Shahriyar Amini, Jason I Hong, Norman Sadeh, Janne Lindqvist, and Joy Zhang. Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 501–510, 2012.

[34] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi, Shikun (Aerin) Zhang, Norman Sadeh, Yuvraj Agarwal, and Alessandro Acquisti. Follow my recommendations: A personalized privacy assistant for mobile app permissions. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 27–41, Denver, CO, June 2016. USENIX Association.

[35] Christina Low, Emma McCamey, Cole Gleason, Patrick Carrington, Jeffrey P. Bigham, and Amy Pavel. Twitter A11y: A Browser Extension to Make Twitter Images Accessible. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–12, New York, NY, USA, 2020. Association for Computing Machinery.

[36] Aleecia M McDonald and Lorrie Faith Cranor. The cost of reading privacy policies. *Isjlp*, 4:543, 2008.

[37] Andrey Meshkov. "Big Star Labs" Spyware Campaign Affects Over 11,000,000 People. Retrieved May 25, 2021 from https://adguard.com/en/blog/big-star-labs-spyware.html.

[38] Liz Mineo. On internet privacy, be very afraid. *Harvard Gazette*, 2017.

[39] Extension Monitor. Breaking Down the Chrome Web Store. Retrieved September 3, 2020 from https://extensionmonitor.com/blog/breaking-down-the-chrome-web-store-part-1.

[40] Vivian Genaro Motti and Kelly Caine. Users' privacy concerns about wearables. In Michael Brenner, Nicolas Christin, Benjamin Johnson, and Kurt Rohloff, editors, *Financial Cryptography and Data Security*, pages 231–244, Berlin, Heidelberg, 2015. Springer Berlin Heidelberg.

[41] Stefan Palan and Christian Schitter. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

[42] Stephanie E Perrin. *The personal information protection and electronic documents act: An annotated guide*. Irwin Law, 2001.

[43] Qualtrics. Qualtrics, 2020. Retrieved September 4, 2020 from https://www.qualtrics.com.

[44] Marshall David Rice and Ekaterina Bogdanov. Privacy in doubt: An empirical investigation of canadians' knowledge of corporate data collection and usage practices. *Canadian Journal of Administrative Sciences / Revue Canadienne des Sciences de l'Administration*, 36(2):163–176, 2019.

[45] Judy Robertson and Maurits Kaptein. *Modern Statistical Methods for HCI*. Springer Publishing Company, Incorporated, Switzerland, 1st edition, 2016.

[46] Jerome H Saltzer and Michael D Schroeder. The protection of information in computer systems. *Proceedings of the IEEE*, 63(9):1278–1308, 1975.

[47] Stefan Schneegass, Romina Poguntke, and Tonja Machulla. Understanding the impact of information representation on willingness to share information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–6, New York, NY, USA, 2019. Association for Computing Machinery.

[48] Irina Shklovski, Scott D. Mainwaring, Halla Hrund Skúladóttir, and Höskuldur Borgthorsson. Leakiness and creepiness in app space: Perceptions of privacy and mobile app use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 2347–2356, New York, NY, USA, 2014. Association for Computing Machinery.

[49] Oleksii Starov and Nick Nikiforakis. Extended Tracking Powers: Measuring the Privacy Diffusion Enabled by Browser Extensions. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1481–1490, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

[50] StatCounter. Global market share held by leading desktop internet browsers from january 2015 to june 2020 [graph]. *Statista*, 2020.

[51] Gaurav Varshney, Manoj Misra, and Pradeep K. Atrey. Detecting Spying and Fraud Browser Extensions: Short Paper. In *Proceedings of the 2017 on Multimedia Privacy and Security*, MPS '17, pages 45–52, New York, NY, USA, 2017. Association for Computing Machinery.

[52] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[53] Jofish Williams, Alex C. and Cambre, Julia and Bicking, Ian and Wallin, Abraham and Tsai, Janice and Kaye. Toward Voice-Assisted Browsers : A Preliminary Study with Firefox Voice. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, CUI '20, New York, NY, USA, 2020. Association for Computing Machinery.

[54] Yuxi Wu, Panya Gupta, Miranda Wei, Yasemin Acar, Sascha Fahl, and Blase Ur. Your Secrets Are Safe: How Browsers' Explanations Impact Misconceptions About Private Browsing Mode. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, pages 217–226, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[55] Xinyu Xing, Wei Meng, Byoungyoung Lee, Udi Weinsberg, Anmol Sheth, Roberto Perdisci, and Wenke Lee. Understanding Malvertising Through Ad-Injecting Browser Extensions. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, pages 1286–1295, Republic and Canton of Geneva, CHE, 2015. International World Wide Web Conferences Steering Committee.

[56] Liu Yang, Nader Boushehrinejadmoradi, Pallab Roy, Vinod Ganapathy, and Liviu Iftode. Short paper: enhancing users' comprehension of android permissions. In *Proceedings of the second ACM workshop on Security and privacy in smartphones and mobile devices*, pages 21–26, 2012.

A Survey

Demographics

Q1.1 Please enter your Prolific ID here

Q1.2 What is your age?

Q1.3 What is your gender?

Options: Male; Female; Diverse; Other; Prefer not to say

Q1.4 What is your highest level of education?

Options: No formal education; High school diploma or equivalent; Bachelor's degree or equivalent; Master's degree or equivalent; Doctoral degree or equivalent

Q1.5 Are you majoring in or have a degree or job in computer science, computer engineering, information technology, or a related field?

Options: Yes; No

Behaviour and knowledge

Q3.1 Which is your default desktop browser?

Options: Chrome; Firefox; Safari; Edge; Opera; Other (Please specify)

Q3.2 Please select all the desktop browsers that you use to some extent.

Options: Chrome; Firefox; Safari; Edge; Opera; Other (Please specify)

Q3.3 Do you use browser extensions?

Options: Yes; No

Q3.4 (Shown if Q3.3 = Yes) Which type of browser extension do you use? (Select all that apply)

Options:

Advertisement, cookies, or tracker blocker (e.g. Ad-block plus, uBlock origin);

Password manager (e.g. Lastpass, 1Password); Shopping assistant (e.g. Honey, Piggy);

Language tool (e.g. Oxford dictionary, Grammarly);

Productivity (e.g. Todoist, Evernote);

Video or music downloader (e.g. Youtube Downloader, Video DownloadHelper);

I don't use any browser extension;

Other (Please specify)

Q3.5 (Shown if Q3.3 = Yes) Please indicate on the scale; how confident you are that:

Columns: 1 - Not at all confident; 2; 3; 4; 5 - Very confident

Rows:

R1 You know what type of data is collected by your browser extensions;

R2 You know how your data is used by your browser extensions;

R3 Please choose the fourth option;

R4 The developers of your default browser have made sure your data is safe from being tampered with or shared without your consent;

R5 The developers of your browser extensions have made sure your data is safe from being tampered with or shared without your consent

Q3.6 (Shown if Q3.3 = Yes) Please respond to the following questions, in relation to your browser extension:

Columns: Yes; No; I don't remember

Rows:

R1 Have you ever read the privacy policy for any of your browser extensions?;

R2 Have you ever read the terms and conditions for any of your browser extensions?;

R3 Have you taken steps to ensure your data is secure and private for your browser extensions?

Q3.7 (Shown if Q3.3 = No) Please indicate on the scale; how confident you are that:

Columns: 1 - Not at all confident; 2; 3; 4; 5 - Very confident

Rows:

R1 You know what type of data is collected by browser extensions;

R2 You know how user data is used by browser extensions;

R3 Please choose the fourth option;

R4 The developers of browsers have made sure user data is safe from being tampered with or shared without user's consent;

R5 The developers of browser extensions have made sure user data is safe from being tampered with or shared without user's consent.

Q3.8 (Shown if Q3.3 = No) Please respond to the following questions, in relation to browser extensions:

Columns: Yes; No; I don't remember

Rows:

R1 Have you ever read the privacy policy of any browser extension?;

R2 Have you ever read the terms and conditions of any browser extension?;

R3 Have you taken steps to ensure your data is secure and private for any browser extension?

Q3.9 Please indicate on the scale:

Columns: 1- Not at all interested; 2; 3; 4; 5 - Extremely interested

Rows:

R1 Your degree of interest in seeking out information about security and privacy in relation to browser extensions.

Q3.10 Please indicate on the scale; how comfortable you are with:

Columns: 1 - Not at all comfortable; 2; 3; 4; 5 - Extremely comfortable

Rows:

R1 Having everything you do in the browser collected and stored by a browser extension.

Q3.11 (Shown if Q3.3 = No) Why don't you use browser extensions?

Options (randomised):

I don't need them;

I didn't know they exist;

Due to concerns about data privacy;

It's too difficult to install them;

Other:

Q3.12 Assuming that you have an Ad-blocker installed as a browser extension; can it read passwords that you use on various websites?

Options: Yes; No; I don't know

General Scenarios

Q4.1 Please indicate if you think it is technically possible for a browser extension to cause the following scenarios. Also indicate how likely you think the scenario will be used in a malicious way. An installed browser extension:

Columns: G1 Possible (SG1 Yes; SG2 No; SG3 I don't know); G2 Likely to be used in a malicious way (SG4 Very unlikely; SG5 Unlikely; SG6 Neither likely nor unlikely; SG7 Likely; SG8 Very likely)

Rows (randomised):

R1 Reads the user's usernames and passwords and stores them on an external server;

R2 Replaces the product link to e-commerce websites such as Amazon and eBay with an affiliate link;

R3 Replaces the advertisement on the website with advertisement from its own ad network;

R4 Accesses the user's camera and microphone and records a video;

R5 Uninstalls another browser extension;

R6 Installs an application on the user's computer;

R7 Blocks access to a webpage;

R8 Changes the password of the user's social media account;

R9 Restarts the computer;

R10 Changes the default password for the computer.

Specific Scenarios

(Each participants is shown one question out of Q5.1-3 at random)

Q5.1 (Chrome permission dialogue) Given the dialogue below; please indicate if you think it is technically possible for a browser extension, asking for these permissions, to cause the following scenarios. Also indicate how likely you think the scenario will be used in a malicious way. The browser extension:

Columns: G1 Possible (SG1 Yes; SG2 No; SG3 I don't know); G2 Likely to be used in a malicious way (SG4 Very unlikely; SG5 Unlikely; SG6 Neither likely nor unlikely; SG7 Likely; SG8 Very likely)

Rows - Same as in Q4.1 with the randomised order maintained

Q5.2 (Safari permission dialogue) - Same as Q5.1 in other aspects and the randomised order maintained from Q4.1 in rows

Q5.3 (Firefox permission dialogue) - Same as Q5.1 in other aspects and the randomised order maintained from Q4.1 in rows

Analysis of permission statements

Q6.1 Statement A: "The browser extension can access; meaning read and change; all information including sensitive information such as passwords, phone numbers, credit card numbers, text and images on all websites such as those for online banking, email service, online shopping, and social media." Compared to Statement A; please indicate ...

... how similar is the information conveyed by the following permissions.

... your preference for the following permissions in place of Statement A.

Columns: G1 Similarity (SG1 Not at all similar; SG2 Somewhat similar; SG3 Extremely similar); G2 Preference (SG4 Not at all preferred; SG5 Somewhat preferred; SG6 Extremely preferred)

Rows (randomised):

R1 Access all your data for all websites;

R2 Read and change all your data on websites you visit;

R3 Web page content: Can read sensitive information on web pages including passwords, phone number and credit cards on all web pages.

Q6.2 Statement B: "The browser extension can read and modify the privacy settings of your browser. These settings control the information the browser makes available to websites, manage the browser's inbuilt password manager, and control the network connections." Compared to Statement B; please indicate ...

... how similar is the information conveyed by the following permissions.

... your preference for the following permissions in place of Statement B.

Columns: G1 Similarity (SG1 Not at all similar; SG2 Somewhat similar; SG3 Extremely similar); G2 Preference (SG4 Not at all preferred; SG5 Somewhat preferred; SG6 Extremely preferred)

Rows (randomised):

R1 Change your privacy-related settings;

R2 Read and modify privacy settings.

Q6.3 Statement C: "The browser extension can display notifications to you. Notifications can be used to inform you about background processes such as a summary of network requests blocked by an Ad-blocker or combine messages from one or more web services." Compared to Statement C; please indicate ...

... how similar is the information conveyed by the following permissions.

... your preference for the following permissions in place of Statement C.

Columns: G1 Similarity (SG1 Not at all similar; SG2 Somewhat similar; SG3 Extremely similar); G2 Preference (SG4 Not at all preferred; SG5 Somewhat preferred; SG6 Extremely preferred)

Rows (randomised):

R1 Display notifications;

R2 Display notifications to you.

Q6.4 Statement D: "The browser extension can access; meaning read and change; your browsing history. Your browsing history contains information including timestamps and number of visits about the websites that you have opened in the past." Compared to Statement D; please indicate ...

... how similar is the information conveyed by the following permissions.

... your preference for the following permissions in place of Statement D.

Columns: G1 Similarity (SG1 Not at all similar; SG2 Somewhat similar; SG3 Extremely similar); G2 Preference (SG4 Not at all preferred; SG5 Somewhat preferred; SG6 Extremely preferred)

Rows (randomised):

R1 Browsing history: Can see when you visit all web pages;

R2 Access browsing history

R3 Read and change your browsing history.

Privacy policy and terms of use

Q7.1 Please indicate on the scale; the likelihood that you will now:

Columns: 1 - Not at all likely; 2; 3; 4; 5 - Extremely likely

Rows (randomised):

R1 Read the privacy policy for your browser extensions;

R2 Read the terms and conditions for your browser extensions;

R3 Take steps to ensure your data is secure and private for your browser extensions

Q7.2 Please indicate on the scale; the likelihood that you will now:

Columns: 1 - Not at all likely; 2; 3; 4; 5 - Extremely likely

Rows (randomised):

R1 Read the privacy policy if you will install a browser extension;

R2 Read the terms and conditions if you will install a browser extension;

R3 Take steps to ensure your data is secure and private if you will install a browser extension.

Q7.3 Please indicate on the scale:

Columns: 1 - Not at all interested; 2; 3; 4; 5 - Extremely interested

Rows:

R1 Your degree of interest in seeking out more information about security and privacy in relation to browser extensions.

Q7.4 Have you ever seen this or a similar permission dialogue?

Options: Yes; No; I don't remember

Q7.5 (Shown if Q7.4 = Yes) Did the permission dialogue influence your decision about installing the browser extension?

Options: Yes; No

Q7.6 (Shown if Q7.4 = Yes) Please explain your answer to the last question.

B Additional Graphs and Tables

as of September 2020 | Selected five extensions for our study

Extension	In Firefox top 10?	In Chrome top 10?	In Edge top 10?	In Opera top 10?	In Safari top 10?	Appears on x top lists across browsers	Number of users/ downloads	Ratings/ Reviews	Number of requested permissions
Chrome									
Adblock - best ad blocker	X	X		x	x	4	+10.0M	295K	2
Adblock Plus	X	X		x		3	+10.0M	171K	2
Honey		X	x			2	+10.0M	158K	1
Adblock for Youtube		X				1	+10.0M	113K	
Google Translate		X				1	+10.0M	43K	
Grammarly for Chrome	X	X			x	3	+10.0M	38K	2
Avast Online Security		X				1	+10.0M	24K	
uBlock Origin	X	X	x	x		4	+10.0M	22K	2
Adobe Acrobat		X				1	+10.0M	11K	
Avast SafePrice		X				1	+10.0M	11K	
Safari									
Magic Lasso Adblock for Safari					x	1		928	
Adblock for Safari	x	x		x	x	4		902	2
Rakuten Ebates Cash Back					x	1		718	
Grammarly for Safari	x	x			x	3		613	2
Unicorn Blocker:Adblock					x	1		346	
Notebook - Take Notes, Sync					x	1		245	
StopTheMadness					x	1		194	
Mate: Universal Tab Translator					x	1		159	
Ka-Block!					x	1		152	
Ecosia					x	1		146	
Firefox									
Adblock Plus	X	x		x		3	6.8M		6
uBlock Origin	X	x	x	x		4	3.8M		6
Easy Screenshot	X					1	3.0M		
Video DownloadHelper	X					1	2.3M		
Cisco Webex Extension	X					1	2.2M		
Facebook Container	X					1	1.5M		
Grammarly for Firefox	X	x			x	3	1.1M		3
DuckDuckGo Privacy Essentials	X					1	1.0M		
Ghostery - Privacy Ad Blocker	X			x		2	1.0M		
Adblock for Firefox	X	x		x	x	4	1.0M		6
Excluded browsers in our study									
Edge									
WindmillVPN - Fast,Safe,Best VPN & Proxy			X			1		708	
G-Translate			X			1		650	
Norton Safe Web			X			1		619	
Honey		x	X			2		530	
uBlock Origin	X	x	X	x		4		522	
AdGuard AdBlocker			X	x		2		514	
Tampermonkey			X			1		428	
Video Downloader professional			X			1		387	
YouTube Video Downloader and MP3 converter			X			1		309	
Hola Free VPN proxy Unblocker - Best VPN			X			1		307	
Opera									
SaveFrom.net helper				x		1	87.2M	3467	
Adblock Plus	x	x		x		3	40,7M	2625	
Adblock	x	x		x	x	4	14.2M	1212	
Install Chrome Extensions				x		1	13.6M	2611	
360 Internet Protection				x		1	8.8M	687	
Adguard			x	x		2	7.9M	2303	
uBlock Origin	x	x	x	x		4	7.1M	1580	
Translator				x		1	5.8M	2063	
Ghostery	x			x		2	5.6M	946	
Amazon for Opera				x		1	5.5M	307	

Table 2: Table of the top 50 most used browser extensions across Chrome, Safari, Firefox, Edge and Opera as of September 2020. In addition, the number of requested permissions are listed for the five browser extensions we selected for our survey study.

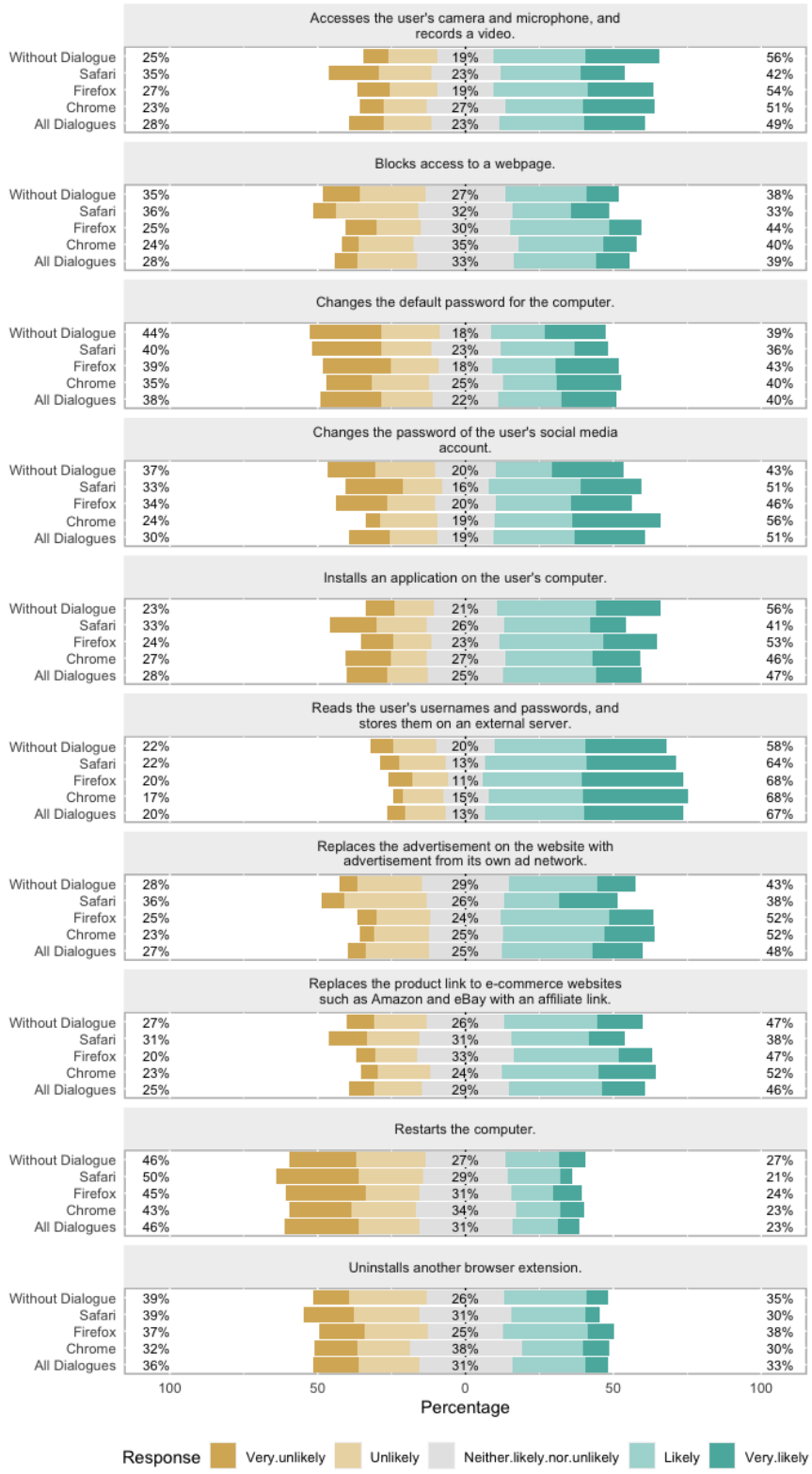


Figure 8: Impact of permission dialogues on participants' perception of the likelihood of scenarios being used maliciously.

Replication: Effects of Media on the Mental Models of Technical Users

Khadija Baig, Elisa Kazan, Kalpana Hundlani, Sana Maqsood, Sonia Chiasson
Carleton University
[*KhadijaBaig, ElisaKazan, KalpanaHundlani, SanaMaqsood*]*@cmail.carleton.ca*
chiasson@scs.carleton.ca

Abstract

Media has been observed to influence users' mental models in several domains. It was recently demonstrated that fictional television and movies have a strong influence on non-technical end users' mental models of security. We extended this study to explore its effect on 23 participants with technical backgrounds, given that misconceptions amongst this group could have important organisational impacts or could influence other non-technical end users. Our qualitative analysis reveals that technical participants sourced their mental models from both their academic or professional lives and from different forms of media (like news, cinema, forums, and social media). They were capable of identifying unrealistic depictions of hacking in the provided video clips and most could offer simplistic explanations about why these were problematic. We found that they generally had more nuanced understanding of the issues than non-technical end users, but they were not immune to misinformation from mass media.

1 Introduction

Users are regularly faced with decisions that impact their security or privacy online. The decisions of individuals in technical roles can impact entire networks, the robustness of software, or trusted advice given to non-technical end users. Many non-technical users look to technical individuals amongst their family, friends, and acquaintances for cybersecurity advice. Incorrect mental models by technical users could directly affect an organisation, and sharing incorrect information could affect the receiver's cybersecurity attitudes and practices. For this reason, technical users need accurate mental models of online security: how computer systems work, methods of protection, and risky behaviours.

Previous studies have shown that media can affect viewers' mental models, having been successfully used as an educational tool in the past (for example, to motivate students to study science [6], or as advertisement campaigns that act as Public Service Announcements [10]). It has also been seen that mental models of online security have been influenced by media in the past [30]. Depictions of cyber-security in media often involve certain tropes: fast-paced, dramatic depictions of hacking, use of technical jargon, decryption that occurs in a span of seconds, and cyber-security attacks mostly happening to large organisations, or individuals with wealth [20, 35, 36].

Incomplete, inconsistent, or inaccurate mental models of cybersecurity can lead non-technical end users to make negative decisions about how they handle their security and privacy online; for example, feeling that SMS and landline phone calls were at least as secure as end-to-end encrypted communication [2]. To understand how non-technical end users evaluate depictions of online security in media, and the effect it has on their existing mental models, Fulton et al. [14] conducted a study with 19 participants of different backgrounds. Fulton's study confirms that non-technical end users often turn to fictional media and its tropes to fill gaps in their technical knowledge. Participants often did not have enough technical knowledge to accurately evaluate a scene and would turn to existing tropes to justify realism; for example, many found technical jargon to be a sign of realism. They also turned to more environmental cues to inform their judgements, evaluating the perceived realism of the situation and characters, and drawing parallels to their own personal experience.

One would assume that users from technical backgrounds would be better informed in this domain, but this is not always the case. Computer Science students and developers alike have been found to have limited understanding of privacy and online security practices [5, 15, 34]. Given that Fulton's study did not control for technical expertise [14], we extend this study with users who have a technical background.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

23 participants took part in our study, which followed the protocol of Fulton et al. exactly. We conducted 60-minute sessions, consisting mainly of an interview supported by video clips of hacking portrayed in popular television shows or movies. In addition to the questions from the original study, we asked participants about their technical background. While there are several studies on looking at users' mental models of the Internet and cybersecurity [3] [40], to the best of our knowledge this is the first study exploring the effect of fictional television and movie media on technical users. We found that technical participants had more complete mental models of hacking and security than non-technical end users. They were generally able to assess the realism of hacking in video clips, but they still had misconceptions, and believed at least some of the inaccurate depictions presented.

2 Background

We discuss existing mental models of security, the effect these mental models have on software security, and the role media plays in information propagation overall.

Mental models of security: Online security is often linked to several digital systems and tools, such as anti-viruses, firewalls, encryption, and web security. When looking at users' mental models of computer security warnings, Bravo-Lillo et al. found that users with greater technical knowledge had more complex mental models than non-expert users [7]. Raja et al. found that users with higher levels of security knowledge often understood the general functionality of a firewall, but were unable to address key parts of its functionality (for example, being unable to identify the effect of choosing a network in their settings) [27]. In a study looking at general mental models of the Internet, it was found that more technical and non-technical users held similar beliefs [19], although technical users did perceive more privacy threats. These beliefs included the idea that attackers only go after high-value targets, and generally are too powerful to be stopped. Assal et al. found that while several developers agreed on the importance of software security, they mostly thought of their applications as not being a worthy target for attackers [4]. These types of beliefs have been found to affect users' security behaviours [18], such as failing to take precautions against broader, non-targeted attacks [14].

Interviews with smartphone app developers reveal concern over the lack of focus on security in technical-related education, with many developers simply turning to the Internet for answers when confronted with such obstacles in their work [5]. Similar sentiments were seen in Tahaei's interviews [34], where Computer Science students did not have holistic perceptions of computer security. These students often drew parallels to Hollywood hacking and cited media as a source for their mental models. Tahaei's study consisted of semi-structured qualitative interviews with Computer Science students, without the use of any external media.

Redmiles et al. surveyed a broad, census-representative US population to shed light on which factors influence users' rejection or adoption of security advice [28]. It is unclear whether their sample includes users with technical backgrounds. They found that the two major sources of online security advice were media and family or friends. 67.5% of respondents cited media as a source, and 60% of the advice given by family or friends were by people with background in Computer Science or IT. With users who received advice at work, more than 50% did so from someone with IT background. The study also found, however, that users with higher internet skill were 32% more likely to use media as a source of advice. Wash and Cooper [38] found that when being trained against phishing, users are more likely to benefit from security advice if provided by a security expert, and from relevant stories if provided by a peer. Given that several users turn to their more technically versed family, friends, and colleagues for advice, it is increasingly important that technical users have a sound understanding of online security lest they propagate inaccurate advice.

Software security: Millions of users have been affected by exploited vulnerabilities in software [13], despite the existence of best practices for incorporating security into the software development life cycle [22] [26]. Companies and developers have been reported logging unencrypted data in applications [9], and storing sensitive information (like passwords) in insecure areas [12], or storing them insecurely (e.g., unhashed or unencrypted)r [9]. Many posit that if developers had better, more complete knowledge of security, developed applications would be more secure as well [25]. To examine whether developers neglect to write secure code due to their mental models, Naiakshina et al. conducted a study examining whether Computer Science students would store passwords in a secure manner [24]. Their results show that none of the students did so without explicit prompting, and often had little understanding of cryptographic APIs. Students justified that if this was code being written for a real application, they would have done so without prompt. The study was repeated with freelance developers, who were hired to write code for what they believed was a startup-company [23]. These participants also mostly wrote insecure code either unless prompted, with several having misconceptions of password storage security and interchangeably using the terms *hashing* and *encryption*.

Role of media: The effect of media on the consumer has been noted in non-security related contexts, such as promoting knowledge of disease and healthcare. For example, Hether et al. found that exposure to breast cancer storylines affect users' attitudes and behaviours' to the illness, with exposure to multiple storylines being more effective than exposure to a single one [17]. Fulton et al. observed a similar effect on users' cybersecurity knowledge, and discussed how certain media events influence mental models of online security [14]. This further influences user behaviour, like whether they ignore obvious security practices based on the belief that there is no

Table 1: Technical experience. Numbers indicate count of participants per category.

Occupation	Student	10
	Project Manager	2
	IT	3
	Web Developer	1
	Software Developer/Engineer	2
	Network Maintenance	1
	UX Designer	1
	Instructor	1
	Retired	1
	Prefer not to answer	1
Programs of study	CS	9
	Engineering	7
	Business	2
	HCI	1
	Applied Science	1
	Project Management	1
	Prefer not to answer	2
Security exposure	None	10
	Work	6
	Study	4
	Study and Work	2
	No answer	1
Cyber-challenges	Completed	5
	No exposure	18

point. Conversely, the study also found that media could have positive effects on mental models if done correctly. We use their study protocol to evaluate how much media affects the mental models of technical users in this domain.

3 Methodology

Our methodology follows that of the Fulton et al. study [14], with extra questions in the post-test questionnaire. The interview script, post-test questionnaire can be viewed in the Appendix. The study was cleared by our Research Ethics Board. We pilot tested the study with an undergraduate Computer Science student who had reasonable knowledge and experience in cybersecurity; no changes were necessary.

3.1 Recruitment and Participants

Participants were recruited[1] via posters placed around our University campus. The study was also posted on a social media page advertising research studies by the university, and on online service-exchange platforms. We also used snowballing techniques. The eligibility criteria were: (1) being at least 18 years of age, (2) being fluent in English, (3) having normal or corrected vision, and (4) having a technical background.

[1]Note: the study was conducted before the COVID-19 pandemic.

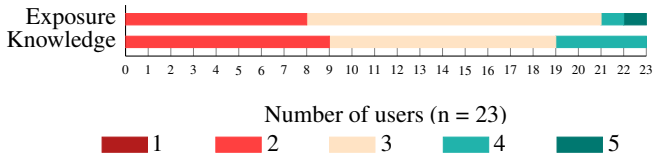


Figure 1: 5-point Likert-scale responses to cybersecurity exposure, knowledge (1 = none; 5 = very high)

We define "technical background" as having academic, work-related, or self-taught exposure to and experience in technical activities. This would include the field of Computer Science, Information Technology (IT), and Systems, Software, and Computer Engineering. It would also include IT and software project managers, freelance, and self-taught programmers. Participants were paid $15.

We initially recruited 26 participants (detailed demographics in Appendix A) but three were excluded due to an error during recruitment. This left us with 23 eligible participants: 11 identified as female and 12 as male. Excluding a participant who preferred not to answer, ages ranged from 18 years to 65 (M = 30.78, SD = 13.09).

All participants had completed post-secondary education. The majority had completed, or were enrolled in, either an undergraduate (n = 16) or a graduate degree/certificate (n = 6). Table 1, and Figure 1 provide more detail on participants' technical backgrounds, including their self-reported exposure (how often they hear about or discuss cybersecurity) and knowledge of cybersecurity (how much they know about cybersecurity). Participants generally reported similar levels of exposure and knowledge, only ever varying by one point.

Participants read and signed a consent form explaining the purpose and procedure for the study, and provided permission to be audio-recorded. Participants were assigned a pseudonym (e.g., P1-S3, P2-S3) that was not linked to their identity. The appended letters indicate participants' self-reported cybersecurity exposure: S1 (None), S2 (A little), S3 (Some), S4 (High), and S5 (Very high).

3.2 Procedure

The study involved completing an online screener questionnaire, followed by either an in-person or remote study session for those who qualified. All questionnaires used in the study were hosted on Qualtrics [2]. The questionnaires had a "prefer not to answer" option available for all questions.

Screener Questionnaire: Prior to being booked for a study session, potential participants completed an online screener questionnaire to assess eligibility. The screener had its own consent form embedded in it.

[2]https://www.qualtrics.com/

Study Session: Qualifying participants were invited to complete a 60-minute session which was audio-recorded with the participant's consent. In-person sessions were completed in our lab. Participants could also complete the study remotely through video-conferencing (e.g., Skype or Google Hangouts). The study session consisted of an interview and a questionnaire. During the interview, participants watched 6 different video clips and evaluated their perceived accuracy.

3.3 Interviews

The interview protocol mirrors the study by Fulton et al. [14], and is as follows:

Mental models: Participants answered questions to help assess their existing mental models of cybersecurity, hacking, and encryption.

Personal experience: Participants described incidents where they, or someone they knew, were being hacked.

Prior media exposure: Participants recalled whether they have seen any, or knew of, fictional TV/movies that contained content related to cybersecurity, hacking, and encryption.

Video clips: Participants watched 6 video clips. At the end of each clip, they answered: (i) whether they could identify the media or the scene, (ii) with a summary of the video clip, (iii) what they found realistic, and (iv) what they found unrealistic.

Realism in media: Participants described their general perceptions of (i) how realistic they find media in portraying these topics, (ii) media that portray these topics realistically, and (iii) media that portray these topics unrealistically.

3.4 Video Clips

We used the same video clips as the original study [14]. These were selected from television programs and movies to cover a wide variety of scenarios, tropes, and levels of realism.While we provide the source of the clip in the following descriptions, participants saw only the video clip without context of where it originated. Videos were played in random order.

1. Superman 3: An employee is disappointed with his first paycheck. A colleague tells him that in every big corporation, there are half-cents left over, but only the computers know where they go. Inspired by this, the employee stays after work and successfully hacks the system by typing in *Override all security*, and then using the command *Reroute all half-cents to above account* to add to his paycheck.

2. The Amazing World of Gumball: A blue and a pink character reach a locked door in a building with a computer terminal next to it. The blue one worries that they can't get in, but the pink one reassures him that she can break through. She types in the letters H-A-C-K, and presses enter. This opens the door, and the blue character is surprised. The pink character reveals she was joking, and explains in very technical terms how she actually hacked the door.

3. NCIS: A forensic team's computer is getting hacked. Several windows pop up and flash on the screen. An agent frantically types on the keyboard while exclaiming "they've broken through the NCIS public firewall!". Another agent joins the first agent in typing on the keyboard (four hands, one keyboard). Suddenly, the screen goes black, and the two are confused as to how the hack stopped, neither believing they were responsible. They look around and notice that another team member had unplugged the computer.

4. Blackhat: a high-level government agent receives an email asking him to change his password. The agent opens a file attached to the email, and downloads a keylogger by doing so. He can now see the new password being typed in by the agent in real time. The hacker then uses these credentials to successfully log in to the system.

5. Sneakers: a blind man sits in front of a computer, and asks another man to name places that are impossible to get into. They start with the federal reserve, and when they bring up its website, its "encrypted", with several nonsensical characters on the screen. Several people watch as the man uses a chip to decrypt everything. He replicates this with the national power grid. He describes encryption as a series of complex mathematical problems that can be broken like any code. Another man states that this chip is *the* code breaker, able to break any kind of encryption.

6. Skyfall: Two men stand in front of a giant computer screen which displays some sort of network that is constantly changing. A laptop is plugged into their computer infrastructure, which they are attempting to hack into. One man notices the name of a station amongst letters flashing on a screen, and asks his team member to use that as a "key". After doing so, the network rearranges to form a map of London. Suddenly several doors open, and they realise they've been hacked. The laptop flashes "Not such a clever boy after all", and while one man runs outside, the other frantically unplugs the laptop from their computer systems.

3.5 Analysis

In total, we recorded approximately 24 hours of audio from the interviews. The first author manually transcribed the interviews, and returned to the audio recordings and interview notes as needed during analysis to add any missing context.

Interview data was analysed using inductive thematic analysis [11]. We iteratively analyzed and created the codebook. Each transcript was reviewed multiple times, ensuring that every transcript was coded with the final codebook.

The first author conducted the interviews and was involved in all stages of the coding process. A second researcher helped code part of the data, and then left the project. A third researcher coded the remaining data and re-coded some

prior data. The detailed process unfolded as follows (see Appendix B). While editing the transcripts, the first author noted any initial themes occurring in the data. These 9 initial themes were then used as the basis for the first codebook. The first author coded two transcripts with this set of codes, then iteratively revised the themes until no new codes emerged, resulting in a second codebook consisting of 16 high-level themes and 87 sub-codes. This second codebook was used in the initial thematic analysis of the first ten transcripts by the first author.

The first author and a second researcher then coded three new transcripts together, and revised the codebook, resulting in a third codebook. The researchers then re-coded the 3 transcripts individually, as well as 3 additional transcripts, upon which a fourth and final codebook with 17 high-level codes and 95 sub-codes emerged. The percentage agreement between the two researchers for the codes on the overlapping transcripts was 98.67% overall.

A third researcher then continued the analysis process alongside the first author. The first author and the third researcher coded 3 transcripts individually using the fourth codebook, and met to clarify any misconceptions. No further revisions to the codebook were needed. Agreement between the two coders on the overlapping transcripts was 99%. All remaining transcripts were analysed or revised using the fourth codebook: the third researcher independently coded an additional 7 transcripts, and revised the codes for 10 transcripts. The first author also coded a further 3 transcripts.

4 Results

We found no overarching connections between technical backgrounds and mental models. After completing the analysis, we grouped participants based on their technical background (e.g., Computer Science, Project Management, Computer Engineering). We then compared results between group. This process was repeated with security knowledge, and security exposure. No patterns emerged in any of these comparisons. Given our small sample size (n = 23), we did not conduct statistical analysis. We present the results of our qualitative analysis, focusing on the main themes arising from the data.

4.1 Existing Mental Models

We first discuss participants' initial mental models and impression of how hacking is portrayed in the media, as described by participants *before* they viewed the video clips.

4.1.1 Profile of a hacker

Hacker persona: Many participants distinguished between ethical and malicious forms of hacking, acknowledging that a hacker could be hired by a company to ensure that their networks remain secure instead of having malicious intentions. Participants' main focus, however, was on those with malicious intentions.

Rather than having their own agenda, malicious hackers were believed to have been hired by others. Some participants explained that malicious data breaches were committed mostly by individuals who had been hired by an external organisation or a national agency, although they were unclear how this process unfolded. For example, P25-S2 expresses: *"a lot of people say they come from organisations overseas. I'm not sure how they'd find each other though; maybe through networks and contacts"*.

Hackers as individuals are mostly seen in a negative light. These were intelligent persons who had malicious intent, who were misguided, or who were suffering from psychological or social challenges. Hackers were described as users who enjoy the challenge of "code cracking" and problem-solving. Some users identified them as misfits or "outcasts" (P1-S3), suggesting they have "a psychological problem" (P22-S3), have "graduated from other forms of crime" (P19-S4), or have malicious intent. A few participants categorised hackers as thrill seekers, or power hungry, perhaps with the intent of creating a *"legacy that withstands time"* (P22-S3). Hackers might also simply be looking for a sense of community, perhaps in an attempt to fit in with a current friend group, or in search of a support network. P22-S3 distinguished between hackers who are "certified" and those who are self-taught, saying *"people who are self-taught are more dangerous; the intention to learn hacking is to go hack someone, otherwise I don't see the need for it"*.

A few participants viewed hackers with a sense of admiration. Hackers were individuals with tremendous skill. They were acknowledged as *"brilliant"* (P6-S3), *"talented"* (P6-S3, P10-S3) and *"intelligent programmers"*(P9-S3). This could possibly be justified based on certain characteristics of hacking seen in the media, like the speed or simplicity with which hacking occurs. As P14-S3 states: *"hackers are intelligent people, and their number of tries (to hack in) is probably way less than normal people. They can do things you can't even imagine"*.

Considering its general portrayal in media, some participants viewed hacking as a storytelling device that the media uses to convey to the audience that the character is *" smart or technical"* (P22-S3). Others expressed doubt over their portrayal: *"if it was the (hacker's) first try (hacking in)... no one's that much of a genius"* (P10-S3).

Hacker motives: Participants acknowledged money or information as a hacker's primary motives, and identified organisations (private and governmental) as the main intended targets since these were viewed as leading to higher rewards. P4-S3 also highlighted how newer companies or systems are more likely to get hacked: *"They haven't been around long so maybe they don't know who they need to be wary of"*. Hacking was viewed as a threat that organisations could learn to avoid with experience and attention. Three participants

considered individual persons to be secondary victims of an organisational breach: while not a direct target, the breached organisation's employees and customers were ultimately personally affected by breaches as well.

In cases where individual persons were noted as targets, three reasons were considered by participants. First, the target was a high-profile individual (n = 6) whose data could bring monetary or reputational rewards to the hacker. Second, participants thought that attacks could occur against random individuals (n = 4) and that the victim was simply the unlucky recipient of misfortune. And finally, participants noted that data breaches could be done for the purpose of *'stalking'* (P1-S3), and that this could occur towards someone the hacker personally knew, or towards an unfamiliar person if the hacker simply found a way to follow the victim's *'routine'*(P5-S5). Some level of victim blaming was apparent. Many felt that inexperienced users of the Internet, or those simply gullible by nature, would be targeted. As a participant explains: *"It's done to whoever seems most accessible... someone who signs up to a lot of things"*(P2-S3).

4.1.2 Human factors

When asked what makes someone an *easy* target for a hack (as opposed to *intended* target), all responses related to user behaviour or human factors instead of characteristics of the technology being used. More than half of participants (n = 13) expect *vulnerable* users to be the easiest targets. Vulnerability was often linked to inexperience, either due to age, general inexperience with technology, or general gullibility.

P5-S5 highlighted the effectiveness of social engineering for phishing, suggesting that anyone could become vulnerable under certain circumstances: *"Tired people who want to relax at the end of the day are more susceptible. People who multi-task and just want shortcuts could also gloss over a moment that could make them slip"*. With this quote, we note an underlying belief that the victim could have prevented the attack had they been more careful, partially holding the victim accountable for the attack. We found that this belief was pervasive and participants cited various security *behaviours* that could make users targets, such as having bad password habits (n = 7) (e.g., easy passwords, repeated use of passwords), not using certain security tools (n = 7) (specifically antivirus software, firewalls), and browsing the Internet carelessly (n = 6) (e.g., accessing sensitive information over public WiFi, visiting *"unhealthy"* (P10-S3) websites, not being careful when clicking on content online).

Most users who mentioned the victim's age as a factor considered the elderly to be most at risk. However, it appears that participants considered users at both extremes to be particularly vulnerable to hackers. P26-S2 mentioned "young kids who don't know they're giving away information that could make them easy targets", feeling that it would be "easy for people to trick them into dangerous situations".

4.1.3 Perceived origins of mental models

Many participants (n = 14) cited their academic or employment background as the primary source for their mental models and understanding of hackers. Some participants (n = 2) also cited their experience with cyber-challenges as affecting their perception: *"(Hacking) reminds me of cybersecurity challenges. It's just problem-solving; you either crack into it or you don't."* (P3-S3). Participants (n = 9) also did research of their own in cybersecurity (like reading articles online), and voluntarily engaged with others (family, colleagues, friends) on the topic: *"I've talked to people who are very interested in these kinds of things, and try to mimic being in the mind of a hacker"* (P5-S5).

Despite saying that media was generally an unreliable source of information, participants recognized that media played a significant role in forming of their mental models of hacking. Fourteen participants identified that their perceptions came from some form of media. Fictional TV or movies (n = 8) were noted as an important source of information. Online sources such as blogs, forums, Social Networking Sites (SNS), and YouTube (n = 14) were also mentioned. One participant explained: *"I watch horror stories on Youtube, and there's usually a hacker in there; that's where my perceptions of the deep web came from"* (P9-S3). News reports about *'data breaches'* or *'identity theft'* (n = 9) also commonly informed participants' mental models of cybersecurity and of hacking.

We noticed how these sources may have informed participants' responses, even prior to them watching the videos. For example, several participants identified a trope in crime/spy-based media, where there is often a technical person on the cast who responds to, or conducts, hacking. While participants generally said they believed these tropes to be inaccurate, their influence was suggested throughout the interviews. For example, when asked about a hacker's goal, P26-S2 responded *"I'm really into [detective show]: maybe in a hostage situation, you'd hack people to use their information to get people to act a certain way"*.

4.2 Characteristics of Realistic Media

In the second half of the interview, participants viewed each video clip and framed their responses with respect to the clips.

More than half of participants (n = 14) found the video clips to be heavily inaccurate. However, the clip from Blackhat was described as *"refreshing"* (P7-S2) by participants for its relatively accurate depiction of phishing. Participants' overall evaluation of each clip mostly hinged on whether the hack or defence seemed realistic. To supplement their evaluation, participants also used contextual and cinematic cues to assess realism. Several participants were critical of the speed and simplicity of the hacks, but some acknowledged that these aspects were probably *"dramatised for the audience"*.

4.2.1 Unplugging could happen but might not work

Participants' responses to unplugging the computer as a possible defence against hacking were dependent on the video and the participants' assumptions about the attack.

With the NCIS video, some reasoned that unplugging a single PC from its power source was ineffective in protecting against a network attack unless the network was disconnected as well. Others thought it was realistic, citing the commonly heard advice *"try turning it off and on again"*(P18-S2, P26-S2). A few participants clarified that this might only work if the hacker had not reached the network, while others explained that it was realistic in the context of that one specific system: *"when the system is off, how can someone hack into (it) if you're not connected to the Internet anymore?"* (P22-S3). P25-S2 observed that *"only (one) computer was being hacked, interestingly enough"*. Additionally, two participants found it unrealistic that unplugging a desktop computer would actually stop it from running. Like a laptop, they expected a backup battery to keep it running.

A similar scene was present in Skyfall, in which a system is hacked into while connected to a laptop. On realising they've been hacked, an actor unplugs the main system from the malicious laptop. Participants recognised that hackers could *"do a lot more if plugged in versus if not"* (P24-S2), and so unplugging in this case would be a *"a good move"*. However, a few were cynical of this action, unsure if *"(they'd) be able to stop it by that point"* (P25-S2).

Although no consensus was reached, participants' technical knowledge enabled them to assess each situation, reason about the conditions under which the attack may be plausible, and determine the extent of realism for themselves.

4.2.2 Hacker and victim profiles must fit

Hackers: Participants often used the characters' physical traits or personality when assessing realism. Participants relied on their own pre-conceived ideas of a typical hacker and found unrealistic any depictions that did not match their imagined hacker. In Skyfall, one character was referred to as *"a programmer dude"* (P3-S3) because he wore glasses and a sweater vest. In Blackhat, it was perceived as unrealistic that such *"attractive"* (P7-S2) and *"decent-looking"* (P6-S3) individuals could be hackers, and several found Gumball to be unrealistic because it was a cartoon, and because a little girl was a hacker. One participant, however, expected hackers to be younger individuals because they are more comfortable with technology. For example, when the main character in Gumball opens a door with the letters H-A-C-K, P6-S3 agreed that *"children can hack (in), so easily"*.

Security clearance: It was commonly believed that breaking through a system's security measures was difficult, so participants found scenes unbelievable if they perceived that the character wouldn't have adequate security clearance. For example, when Richard Pryor overrides all security access in Superman 3, participants were either suspicious or trusting based on their interpretation of his character. Some found it plausible that he would *"know the vulnerabilities"* (P20-S3) of the organisation simply by virtue of him being part of it. Others assumed he was part of the IT department ("he seems to be a programmer" (P3-S3)), and so accepted that he had some level of access. Several, however, didn't believe that *"the main character is smart enough to hack into the system and get the money out"* (P10-S3). Participants' mental models included some organisational understanding of who would have security clearance or administrator privileges, based on their own experiences, and used this practical knowledge in assessing the realism of the video clips.

Behavioural attacks: Participants were less forgiving of a victim who fell prey to phishing, especially one with high-level access as seen in Blackhat. Many believed that the scene was unrealistic because someone with that level of access would *"know better"*, and believed that anyone in that position would have received formal training addressing this topic. Some recalled their own experiences of having received similar training in the workplace. Interestingly, this contradicts participants' earlier explanation that anyone could be vulnerable due to inattention. This indicates that in contrast to non-technical end users, participants expect those handling sensitive information to not be susceptible to attacks leveraging human factors.

Participants thought that phishing attacks were unlikely to succeed in high-security organisations. Participants agreed that getting phishing emails was *"common"* (P3-S3, P16-S3) in the workplace, but explained that it *"shouldn't be that easy to get into someone's computer"* (P24-S2) and that other forms of security would separate the hacker from the system. Participant noted that security tools like firewalls and multi-factor authentication should hinder access by an attacker. Participants' technical background increased their skepticism in attacks that appeared too simple to be realistic.

4.2.3 Setting for the scene must be realistic

Situational context: Participants found cybersecurity events more realistic, when they matched the context of the organization in which they took place. Using Superman 3 as an example, P10-S3 explained: *"every company has this kind of situation where someone can hack in. A person is knowledgeable about Computer Science or networks, or does the payroll, and is knowledgeable about how to hack into other people's systems"*.

Exaggerations or obvious security lapses within the scenes triggered skepticism. P18-S2 conversely noted about Superman 3:*" I'm unsure how he's able to override the system, but the smaller scale seems more realistic"*. Similarly, in Gumball, some participants found it unrealistic that a door of high-value would be left unguarded.

Organisational values: Having the presence of several

people working together was seen as more realistic than watching a lone-wolf breach systems. For example, when watching Sneakers, P6-S3 explained *"people together trying to find a solution is realistic"*. P6-S3 was also critical of the field agent escaping at the end of the Skyfall clip, stating that no *"real leader"* would leave after a system breach. In participants' view, hacking and administering security were collaborative efforts where knowledge and responsibility was collectively shared among several individuals.

Some also found the lack of protocols in response to a system breach to be unrealistic. As P10-S3 stated: *"there are protocols or policies that need to be followed by leaders instead of running out and leaving the audience to imagine their own thing"*. This principle also applied to Blackhat, where a participant mentioned that a high-ranking organisation would probably have multi-factor authentication available for their systems. Again, participants had expectations with respect to how organisations handle security and breaches, and these were informed by their previous experiences or knowledge of how things "should be".

Timeline: Scenes set, or filmed, in the past were judged differently than those set in the present day. For example, P3-S3 found the decryption chip used in Sneakers to be unrealistic for its time because it was unlikely that this type of technology was available then. While watching the Superman 3 clip, some participants said it would be realistic for the company's systems to be insecure since *"they didn't care about cybersecurity back then"* (P7-S2). However, participants had limits to their allowances. Two participants explained that it was unlikely that systems were *"ever that unsafe"* (P15-S3).

4.2.4 Hacking is stealthy, malware is obvious

The distinction between malware and hacking was somewhat blurred by participants. For example, participants were divided on the accuracy of the NCIS scene where the system displayed several pop-ups after getting hacked. Most participants suggested that this appeared to be malware; their mental model of the association between malware and hacking determined their evaluation of the clip. If participants thought that malware equated to hacking, then pop-ups were to be expected. If participants thought hacking was distinct from malware, then the clip was a clear exaggeration.

Malware: Participants who identified a link between adware (and other malicious software) and hacking recognized the pop-ups, comparing it to their own experiences of visiting a *"bad website"* (P18-S2) or clicking a suspicious link. These participants expected obvious signs that the computer was being hacked: *"I think... malware tried to disrupt [the system]. I've never experienced this before, but I think this is what happens when a system's being hacked. It's a very astonishing thing when a system is being attacked"* (P13-S3). Another participant explained *"I can imagine, in reality, many things popping up as a system is being hacked, many things being stolen. "* (P16-S3). In these cases, the clip reinforced their (mis)understanding of how hacking typically occurs.

Stealth: Other participants were skeptical about whether a hacked computer would *"go that ham"* (P1-S3), possibly because hackers *"want to be undetected "* (P1-S3). Six participants were highly critical of the idea that a hack would have any visible effect, even if done via malware. Hackers would want to go unseen to avoid alerting the user while completing their task of interest. As P7-S2 states: *" Why would a hacker create code that would do that? To tip off the person being hacked?"*. In comparison, the Skyfall clip was more believable for this group of participants. In the clip, the hack was subtle and stealthy; the main characters mostly had no idea that they were being hacked.

Additionally, a few participants (n = 3) did not expect certain aspects of hacking to be *"broadcast to the public"* (P13-S3), for fear it might be *"dangerous"* (P13-S3). For example, a powerful decryption chip would be *"well hidden"* (P24-S2) if it existed, possibly only used by *"high profile (individuals and organisations) and underground cartels"* (P24-S2).

4.2.5 Hacking is complex

Participants expect breaching a system to be a complex process, and not just possible with *"one key-stroke"* (P3-S3). However, three participants noted an exception to this rule: breaching a system can be quick and simple if you have contacts or work for the organisation and are familiar with its vulnerabilities; in other words, insider attacks can be simple. Having a relationship with a *"higher-up"* (P10-S3) or using bribery to obtain information (P10-S3, P13-S3) would provide access to a system through relatively official channels, without need for complexity.

Participants believed organisations to have multiple layers of security in place for their systems. As such, a hacker would probably have to breach several protocols to successfully hack a system using conventional means. While Gumball was praised for its depiction by some participants, others were confused about how the pink character had enough time to truly hack in to the system while on screen. Regardless of these differences in perception, many agreed that the *"long list of things"* that needed to happen, as described by the character, were plausible.

The majority of our participants (n = 15) trusted encryption to be secure. When faced with a chip that destroyed all encryption simultaneously in Sneakers, many found it to be unrealistic. This was partially due to the hardware that would be required, with some feeling that this chip might only be possible *"in the future, maybe"* (P6-S3) and does not *"exist on this planet right now"* (P10-S3, P15-S3). One participant was entirely unconvinced: *"the amount of math you'd have to do would be... wow... in even existing or future computer hardware"* (P7-S2).

As part of a system's defence, participants expect a realistic

system to detect and log any attempted breaches. They criticised clips that showed hackers getting through undetected. Participants believe that most hacking attempts would be unsuccessful in real life. They further expect systems to flag most unauthorised attempts immediately and that these alerts would be immediately actionable by system administrators.

Two participants expressed that since not all organisations and businesses prioritise cybersecurity, some systems may actually be as simple to hack as portrayed. For example, after watching the scene in Gumball where typing H-A-C-K unlocks a door, P11-S2 expressed: *"I think a lot of places put very basic, easy passwords that are very easily guessed"*.

4.2.6 Media hacking is exaggerated, dramatised

Almost all participants (n = 21) commented on the cinematography and artistic liberties taken in the clips during their interview. Participants often expressed their disdain for cinematics, agreeing that while certain depictions *"make for good TV"* (P7-S2), they are often *"exaggerated"* (P9-S3). This includes situational context; As P24-S2 explained, they found the technology used by hackers in Blackhat to be *"too high-tech"* in comparison to the computer used by the victim of the hack, who was a government official. Other examples included the frantic hammering of keys on a keyboard, and random snippets of code flashing on the screen.

Interestingly, only five participants explicitly mentioned that two people were typing on the same keyboard in NCIS. These participants were either amused *"I don't think they're so close that they can finish each others' sentences"* (P3-S3), or instantly dismissed it as unrealistic *"Yeah, no. That's not going to work"* (P4-S3). A few others (n = 3) dismissed the entire clip as unrealistic.

4.3 Evaluating realism

Participants relied on their past experiences and their technical knowledge to assess realism.

Participants were quick to comment if they found certain parts of a video relatable. Some would draw on their technical experience to elaborate how they had *"seen this happen before"* (P3-S3, P7-S2) in the real world (either personally, or to someone else on the news). This was especially obvious with the scene involving a keylogger; 9 participants referred to real world examples. P22-S3 explained *"When you go to online support, they send you a file, and you install it and they are able to move your cursor for you, so it's possible"* (P22-S3).

Participants also used their technical experiences to dismiss certain scenes. For example, after watching the Skyfall video clip, P26-S2 commented *"I haven't experienced a program that has, once it's realised it's gotten hacked, that has a fail-safe measure to hack the hacker"*. When something technical is happening on the screen but participants don't understand it, many were confident enough to deem it unrealistic. The Blackhat clip includes a character dragging and dropping something towards the end of the scene. Participants who noticed this had *"no idea what was going on there"* (P3-S3), and responded by rejecting the premise entirely: *"It was confusing that they dragged and then a bunch of things happened on the screen. I don't know what kind of system does that stuff, so it's not too realistic."* (P4-S3). Others dismissed a scene because they became suspicious when they couldn't make sense of the technical jargon. Using Gumball as an example, P3-S3 stated: *"some stuff she said didn't seem right. It didn't connect, it just seemed like a list of things"*.

As we interviewed participants with a range of technical skills, we found that not all had the same level of cybersecurity knowledge, understanding, or past experiences. Some were able to use their past experiences and knowledge to correctly interpret the information presented in the video clips.

Others, however, were unable to do so. For example, when asked about which mechanisms hackers apply, some struggled to identify any methods beyond phishing, only stating that hackers somehow gain access to a network. When these participants were unsure of the technical nuances in a clip, they relied on their existing technical knowledge to assess its credibility, which was inadequate. Participants attempted to fill in the gaps whenever they were unsure of what was happening in the scene. This resulted in subjective interpretation of cues like technical jargon. Some participants chose to ultimately trust a scene, declaring: *"it could happen in real life"* (P13-S3); others were less committal in their phrasing, accepting that *"they seem to know what they're doing"* (P2-S3).

Others relied on their knowledge to make assumptions about feasibility. In one scene, P26-S2 felt that the *"the coding doesn't seem realistic"*. The participant then provided a counter-example: *"In The Matrix, the fact that they used binary is more realistic and the computer would understand it as opposed to human sentences"* (P26-S2). Similarly, in the Blackhat clip, a keylogger is downloaded onto an individual's computer when they click and open a PDF file. Some participants doubted this transmission vector and, as such, dismissed this threat. Similarly questionable claims were made on the topics of encryption, authentication, antivirus, and firewalls.

5 Discussion

Our study exploring the technical users' perceptions of cybersecurity in media resulted in three main findings:

1. Our technical participants appeared to have a semi-reasonable ability to assess the realism of hacking scenes. Due to their technical knowledge, they had more detailed background understanding, which they used to assess the realism of the clips. Specifically, their articulated reasons for why something was unrealistic were more detailed than those observed in Fulton et al.'s [14] study. However, an occasional gap in their mental models sometimes led them

Table 2: Comparison of non-technical end-users (from Fulton et al. [14]) and technical users' mental models. Section numbers for the associated results are provided for reference.

Topic	Non-technical users	Technical users
Unplugging	Unplugging the computer stops the hacker	[§4.2.1] Unplugging might stop the hacker, but it is more likely if unplugging from the network rather than the power source.
Detectability	Attacks and unsafe situations are obvious	[§4.2.4] **Malware** is what causes obvious pop-ups; hackers probably want to remain undetected
Encryption	Encryption is fragile and all security measures are futile	[§4.2.5] Encryption is nearly impossible to circumvent and security measures can be effective if used appropriately
Targets	Hackers have specific, important targets	[§4.1.1] Hackers have general financial or information goals and rarely target specific individuals.
Phishing	Users should be careful when evaluating suspicious links	[§4.1.2] Users should be careful, [§4.2.2] especially high profile victims who 'should know better'
Realism	To evaluate realism, non-technical end users use technical and non-technical knowledge, assess plausibility of plot and characters, consider cinematic cues	[§4.3] To evaluate realism, technical users use mostly technical knowledge, assess plausibility of plot, characters, location, context, and cinematic cues
Complexity	If it's too quick or easy, it's unrealistic.	[§4.2.5] Too quick and easy is unrealistic, except in cases of insider threats, organisations with lax security measures, and lax defence

to make inaccurate assumptions, which overlapped with those of non-technical, home computer users [37].

2. We found no consensus amongst participants over which of the characters' or systems' actions were unrealistic, demonstrating high variability in the aspects of a scene which they found believable or questionable. For all clips, at least some participants gave inaccurate explanations despite their technical backgrounds. Some participants had polar opposite impressions about the realism of an action.

3. Our data suggests that participants may also be influenced by media and believed at least some inaccuracies, though it is unclear if media informs or reinforces existing mental models. Despite their technical background, participants were not immune to misinformation.

5.1 Comparison with earlier results

Table 2 summarises our results compared to Fulton et al.'s original study [14]. We discuss the over-arching themes present in the two studies.

Unplugging: Some participants agreed that unplugging a device that's being hacked from its power source may be an effective way to stop a hack. However, many generally found unplugging from a network to be more effective at defending against a hack (unless the hack was local to the machine). This distinction was not present in the original results.

Detectability: non-technical end users believed that they would be able to recognize if a system was being hacked or if they encountered an unsafe situation online; the attack would be apparent to the user who could then take steps to mitigate the issue. Technical users believed that malware could cause pop-ups on the screen, but many believed that this was distinct from hacking. Hackers, they believed, would want to be stealthy so that they remain undetected.

Encryption: non-technical end users believed that encryption could be easily broken by skilled enough individuals. Hackers were seen as having an immense amount of power, encryption was futile because hackers could circumvent it, and the idea that hackers had a key that could decrypt everything seemed plausible. However, this point of view was not shared by our technical participants: some participants did believe hackers to be highly talented individuals, but many believed that encryption was strong and would require immensely powerful hardware to crack. The existence of such a "decryption" chip was placed in the far future.

Targets: When discussing intended targets of a breach, there is overlap between non-technical end users and our technical participant. Specific individuals, national organisations, and private businesses were viewed as plausible targets of attack by both groups. Our participants additionally felt that users and businesses with poor security practices were more susceptible to hackers and more likely to be targeted.

Phishing: Both non-technical end users and technical participants agreed that opening unknown and suspicious emails was a precursor to getting hacked. Many confirmed seeing such emails in their own inboxes, and were familiar with incidents of individuals or organisations being breached by way of phishing. Technical participants, however, placed significant responsibility on the victim in these situations, particularly those they considered "high-profile".

Realism: Methods of assessing realism were mostly con-

sistent between non-technical end users and technical participants. Both groups used personal experience, technical knowledge, context, and cinematic cues to evaluate the plausibility of cybersecurity portrayals in media. They were more likely to judge something as realistic if they had either experienced it themselves, or were familiar with someone who had. Use of technical knowledge in such appraisal, however, differs slightly amongst the two groups: technical participants were slightly more critical of technical jargon, expecting it to make sense, and they may have focused more on system vulnerability than non-technical participants. With respect to cinematics, technical participants focused on the realism of the set and how it fit with the context of the scene. For example, they considered what kind of organisation it portrayed and how the characters interacted with each other. non-technical end users, on the other hand, noted audio cues such as dramatic music in their assessment of realism.

Complexity: Hacking that was portrayed through quick, easy tasks was largely seen as unrealistic by both non-technical end users and technical participants. Technical participants commented that hacking is never that simple, unless it is done with help from a human insider. Our participants expected systems to be heavily defended using multiple protocols, and thus were critical of how hacking was portrayed as easy. This is in stark contrast to the original results: end users believed that hacking was easy but expected defence against it to be difficult.

Overall, we find that our technical participants had more nuanced understanding of hacking and security, based on their technical knowledge than the non-technical end users from the original study [14]. However, technical participants also appeared susceptible, although to a lesser degree, to misunderstandings and to believing that some of the fictional portrayals of hacking were realistic. Even participants who demonstrated reasonable knowledge of computer security concepts would occasionally mention *"you see it in the movies"*(P4-S3) as justification for penning a scene as realistic. The varied results of our study are concerning: our participants currently hold, or will soon hold, employment in technical positions. In these professional roles, they may make decisions regarding network configurations, they may administer systems, they may design and develop software, or they may make other decisions that could impact an organisation's susceptibility to security hazards like hacking or ransomware. They may also be in a position to recognize and act against possible security breaches. In any of these roles, accurate interpretations of hacking are especially important due to the potential consequences of their actions.

Our participants also hold informal roles as advisors, tutors, or troubleshooters of computer-related issues for the non-technical people in their lives. Any misconceptions about hacking held by our participants may get propagated amongst this wider circle of individuals who may not be equipped to counter them. Inconsistent or inaccurate advice could lead to further confusion and gaps in the mental models of non-technical users. Additionally, inaccurate advice that matches what is seen in fiction would reinforce the trust non-technical users place in (mostly inaccurate) media depictions.

As such, it is particularly dangerous for participants in technical fields to hold inaccurate, or conflicting, mental models, as it would not only affect them, but also others on both an individual and organisational scale.

5.2 Recommendations

Several suggestions for addressing these misconceptions have already been offered by Fulton et al. [14] for non-technical end users, and they likely largely apply to technical users too. We discuss these, along with additional recommendations.

R1. Security education: Much like Fulton et al. [14], we found that participants relied mostly on their technical knowledge in assessing realism. Participants also tended to adopt stereotypical beliefs about hackers and tended to 'victim-blame' end-users for security failures. This is problematic because technical users may propagate these attitudes in the workplace or to non-technical users that they advise. In both cases, this can undermine the implementation and maintenance of effective security mechanisms and practices [16, 31] Notably, participants had gaps in their mental models that sometimes led to inaccurate assumptions about encryption, the visibility of hacking attacks, who hackers target, how to mitigate threats, and the identity of hackers. As such, we advocate for a more thorough cybersecurity curriculum that addresses both technical details and human factors, and that explicitly tackles common stereotypes and misconceptions.

Individuals with a technical background may not have education specifically on cybersecurity topics. Others may find themselves in an occupation making technology-related decisions without related formal education(e.g., project management within a software team). As such, we suggest including mandatory cybersecurity education within the general education system or as part of workplace training. Prior research has found that the introduction of cybersecurity curriculum as early as elementary and middle school improves digital literacy and cybersecurity awareness [21, 41, 42]. We also urge the application of security in different platforms and services be included to provide a more holistic education (e.g., Abu-Salma et al. [1] found users' perceptions of private browsing mode to be mostly incorrect, while Wermke et al. [39] found users' security mental models of cloud services to be incomplete and undeveloped).

As mentioned by Fulton et al. [14], educators, designers, and developers who are more familiar with the nuances and depth of misconceptions held by target user groups would be able to better address them in their educational material. More broadly, we advocate for closer integration of cybersecurity content within core Computer Science/Engineering curriculum, so that upcoming generations of technical users have a

foundational understanding of secure computing.

R2. Fact-checking databases: We suggest adding educational information about classic television or movie tropes relating to cybersecurity and popular Hollywood hacking dramatizations to fact-checking websites such as Snopes [32]. In effect, we recommend bringing to the forefront these common misconceptions to make them easy for individuals to identify and to correct. Users are increasingly being taught to identify misinformation and verify the authenticity of online sources; we suggest that misinformation from television and movies be treated similarly (and given the popularity of streaming sites, the differentiation between television, movie, and online content is increasingly blurred). Making this information easily accessible online creates an opportunity to educate users when they specifically seek out the information (e.g., when searching about a particular scene or episode).

R3. Using media to educate: As discussed by Fulton et al. [14], we stress the need for collaboration between the entertainment industry and the cybersecurity community. Specifically, we emphasise using media as a tool to increase awareness of cybersecurity concepts. Research suggests that users ration the amount of effort put into security practices, and that asking them to follow certain existing security advice is unreasonable [33]. However, new security practices do exist that require less cognitive effort [16] (e.g., the use of password managers over traditional password security advice). Studies show that non-expert users' practices have remained largely unchanged, and that expert users also mostly employ these same practices [8] despite better knowledge of "best practices". Individuals may also need regular reminders to effectively retain and apply security information [29]. Therefore, a change might be required for how this advice is imparted to the general public. Media may be key to effectively educating users where previous methods have failed. If utilised, we might make progress in normalizing security best practices. For example, the show *Mr. Robot* is noted for its realism; many participants acknowledged its potential as an educational tool by employing realistic depictions of cybersecurity. It is, however, key that the entertainment industry utilizes reliable sources of cybersecurity expertise (as opposed to more general technical sources), to avoid propagating misconceptions such as those observed in our study.

5.3 Limitations

We have a relatively small sample (n = 23) and focused only on qualitative data. Additionally, participants' self-reported levels of cybersecurity knowledge and exposure may not be accurate, given its subjective nature. Our eligibility criteria may have primed participants to consider their technical background as a source for their mental models. Our interview may also have primed participants to: (i) suggest behavioural factors when asked for what makes someone an easy target, and (ii) explicitly look for unrealistic components within the video clips; it is possible that these same participants could have watched these television shows or movies in another context without even noticing or reflecting on their realism. As we followed the exact study methodology of Fulton et al., we similarly did not inquire about participants' perceptions of the actors in the video clips. Since Fulton et al. did not control for demographics, technical participants may have also been included in their study, so we are unable to assess the extent of overlap between the two populations.

5.4 Future work

It would be interesting to compare these results to users' perceptions of more realistic depictions of cybersecurity. Exploring alternate forms of media that our participants cited would also be helpful: blogs, forums, and videos seen on SNS (like YouTube). It also remains to be investigated whether participants "living" the experience through games that emulate hacking would evaluate their experience as realistic or not. The genre in which cybersecurity incidents are portrayed may also have an effect on users' perceptions of the topic. Finally, due to the qualitative nature of our study, we were unable to analyse whether there were links between participant demographics and how likely they were to source their mental models from media. A larger scale study would aid in answering some of these questions. Follow-up studies could make use of a true/false scheme for analysis, by having participants characterise whether media portrayals are accurate.

6 Conclusion

We conducted interviews with 23 participants with technical background to evaluate the effect of fictional television and movie media on participants' mental models of hacking and computer security. Participants were generally capable of determining the realism of hacking scenes, but gaps in their mental models sometimes lead to inaccurate assumptions. We also observed considerable variability among participants with regards to which actions participants identified as unrealistic and to the interpretation of the scenes. In comparison with the study of non-technical end users completed by Fulton et al., we found that our technical participants generally had a better informed or more nuanced assessment of the realism of the attacks. However, our participants were not immune to believing misinformation about hacking that they had previously seen in mass media.

Acknowledgments

We thank our participants and anonymous reviewers for their shared insight. This research was supported by NSERC Discovery Grant RGPIN 06273-2017; and the Canada Research Chairs program under Grant 950-231002-2016.

References

[1] Ruba Abu-Salma and Benjamin Livshits. Evaluating the end-user experience of private browsing mode. In *The 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[2] Ruba Abu-Salma, Elissa M Redmiles, Blase Ur, and Miranda Wei. Exploring user mental models of end-to-end encrypted communication tools. In *Workshop on Free and Open Communications on the Internet (FOCI)*. USENIX, 2018.

[3] Ruba Abu-Salma, M Angela Sasse, Joseph Bonneau, Anastasia Danilova, Alena Naiakshina, and Matthew Smith. Obstacles to the adoption of secure communication tools. In *Symposium on Security and Privacy (SP)*, pages 137–153. IEEE, 2017.

[4] Hala Assal and Sonia Chiasson. 'Think secure from the beginning': A survey with software developers. In *Conference on Human Factors in Computing Systems (CHI)*, page 289. ACM, 2019.

[5] Rebecca Balebako and Lorrie Cranor. Improving app privacy: Nudging app developers to protect user privacy. *IEEE Security & Privacy*, 12(4):55–58, 2014.

[6] Miri Barak, Tamar Ashkar, and Yehudit J Dori. Learning science via animated movies: Its effect on students' thinking and motivation. *Computers & Education*, 56(3):839–846, 2011.

[7] Cristian Bravo-Lillo, Lorrie Faith Cranor, Julie Downs, and Saranga Komanduri. Bridging the gap in computer security warnings: A mental model approach. *IEEE Security & Privacy*, 9(2):18–26, 2010.

[8] Karoline Busse, Julia Schäfer, and Matthew Smith. Replication: no one can hack my mind revisiting a study on expert and non-expert security practices and advice. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS) 2019)*, 2019.

[9] Pedro Canahuati. Keeping passwords secure. https://about.fb.com/news/2019/03/keeping-passwords-secure/, Last accessed: December 2019.

[10] Barbie Clarke and Catherine Gardner. Concerned children's advertisers leads the way. *Young Consumers*, 2005.

[11] Satu Elo and Helvi Kyngäs. The qualitative content analysis process. *Journal of Advanced Nursing*, 62(1):107–115, 2008.

[12] William Enck, Damien Octeau, Patrick D McDaniel, and Swarat Chaudhuri. A study of Android application security. In *USENIX Security Symposium*, 2011.

[13] Equifax. 2017 Cybersecurity Incident & Important Consumer Information. https://www.equifaxsecurity2017.com/frequently-asked-questions/, Last accessed: December 2019.

[14] Kelsey R Fulton, Rebecca Gelles, Alexandra McKay, Yasmin Abdi, Richard Roberts, and Michelle L Mazurek. The effect of entertainment media on mental models of computer security. In *Symposium on Usable Privacy and Security (SOUPS)*. USENIX, 2019.

[15] Irit Hadar, Tomer Hasson, Oshrat Ayalon, Eran Toch, Michael Birnhack, Sofia Sherman, and Arod Balissa. Privacy by designers: software developers' privacy mindset. *Empirical Software Engineering*, 23(1):259–289, 2018.

[16] Cormac Herley. So long, and no thanks for the externalities: The rational rejection of security advice by users. In *The 2009 Workshop on New Security Paradigms Workshop*, NSPW '09, page 133–144. ACM, 2009.

[17] Heather J Hether, Grace C Huang, Vicki Beck, Sheila T Murphy, and Thomas W Valente. Entertainment-education in a media-saturated environment: Examining the impact of single and multiple exposures to breast cancer storylines on two popular medical dramas. *Journal of health communication*, 13(8):808–823, 2008.

[18] Adele E Howe, Indrajit Ray, Mark Roberts, Malgorzata Urbanska, and Zinta Byrne. The psychology of security for the home computer user. In *Symposium on Security and Privacy*, pages 209–223. IEEE, 2012.

[19] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. "my data just goes everywhere:" user mental models of the internet and implications for privacy and security. In *Symposium On Usable Privacy and Security (SOUPS)*, pages 39–52, 2015.

[20] Rhiannon L. Hacker's Game: 10 things Hollywood got wrong about computer hacking. https://www.hotbot.com/blog/10-things-hollywood-got-wrong-about-computer-hacking/, Last accessed: December 2019.

[21] Sana Maqsood, Christine Mekhail, and Sonia Chiasson. A day in the life of jos: A web-based game to increase children's digital literacy. In *The 17th ACM Conference on Interaction Design and Children*, pages 241–252, 2018.

[22] Microsoft. Microsoft Security Development Lifecycle (SDL). https://www.microsoft.com/en-us/securityengineering/sdl/, Last accessed: December 2019.

[23] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, Emanuel von Zezschwitz, and Matthew Smith. " if you want, i can store the encrypted password": A password-storage field study with freelance developers. In *Conference on Human Factors in Computing Systems (CHI)*, page 140. ACM, 2019.

[24] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, Marco Herzog, Sergej Dechand, and Matthew Smith. Why do developers get password storage wrong?: A qualitative usability study. In *SIGSAC Conference on Computer and Communications Security*, pages 311–328. ACM, 2017.

[25] Duc Cuong Nguyen, Dominik Wermke, Yasemin Acar, Michael Backes, Charles Weir, and Sascha Fahl. A stitch in time: Supporting android developers in writingsecure code. In *The 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1065–1077. ACM, 2017.

[26] Berkeley Information Security Office. Secure coding practice guidelines. https://security.berkeley.edu/secure-coding-practice-guidelines#Secure%20coding%20principles, Last accessed: December 2019.

[27] Fahimeh Raja, Kirstie Hawkey, Pooya Jaferian, Konstantin Beznosov, and Kellogg S Booth. It's too complicated, so iI turned it off!: expectations, perceptions, and misconceptions of personal firewalls. In *Workshop on Assurable and Usable Security Configuration*, pages 53–62. ACM, 2010.

[28] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. How I learned to be secure: a census-representative survey of security advice sources and behavior. In *SIGSAC Conference on Computer and Communications Security*, pages 666–677. ACM, 2016.

[29] Benjamin Reinheimer, Lukas Aldag, Peter Mayer, Mattia Mossano, Reyhan Duezguen, Bettina Lofthouse, Tatiana von Landesberger, and Melanie Volkamer. An investigation of phishing awareness and education over time: When and how to best remind users. In *Symposium on Usable Privacy and Security (SOUPS)*, pages 259–284, 2020.

[30] Scott Ruoti, Tyler Monson, Justin Wu, Daniel Zappala, and Kent Seamons. Weighing context and trade-offs: How suburban adults selected their online security posture. In *Symposium on Usable Privacy and Security (SOUPS)*, pages 211–228, 2017.

[31] Martina Angela Sasse, Sacha Brostoff, and Dirk Weirich. Transforming the 'weakest link'—a human/computer interaction approach to usable and effective security. *BT technology journal*, 19(3):122–131, 2001.

[32] Snopes. Snopes. https://www.snopes.com, Last accessed: February 2021.

[33] Elizabeth Stobert and Robert Biddle. The password life cycle: user behaviour in managing passwords. In *10th Symposium On Usable Privacy and Security (SOUPS) 2014)*, pages 243–255, 2014.

[34] Mohammad Tahaei, Adam Jenkins, Kami Vaniea, and Maria Wolters. "i don't know too much about it": On the security mindsets of computer science students. *Socio-Technical Aspects in Security and Trust (first ed.), Thomas Groß and Tryfonas Theo (Eds.). Springer International Publishing. https://www.springer. com/book/9783030559571*, 2020.

[35] TV tropes. Hollywood encryption. https://tvtropes.org/pmwiki/pmwiki.php/Main/HollywoodEncryption, Last accessed: December 2019.

[36] TV tropes. Hollywood hacking. https://tvtropes.org/pmwiki/pmwiki.php/Main/HollywoodHacking, Last accessed: December 2019.

[37] Rick Wash. Folk models of home computer security. In *The Sixth Symposium on Usable Privacy and Security*, page 11. ACM, 2010.

[38] Rick Wash and Molly M Cooper. Who provides phishing training? facts, stories, and people like me. In *The 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.

[39] Dominik Wermke, Nicolas Huaman, Christian Stransky, Niklas Busch, Yasemin Acar, and Sascha Fahl. Cloudy with a chance of misconceptions: Exploring users' perceptions and expectations of security and privacy in cloud office suites. In *Symposium on Usable Privacy and Security (SOUPS)*, pages 359–377, 2020.

[40] Justin Wu and Daniel Zappala. When is a tree really a truck? Exploring mental models of encryption. In *Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 395–409, 2018.

[41] Leah Zhang-Kennedy, Yomna Abdelaziz, and Sonia Chiasson. Cyberheroes: The design and evaluation of an interactive ebook to educate children about online privacy. *International Journal of Child-Computer Interaction*, 13:10–18, 2017.

[42] Jun Zhao, Ge Wang, Carys Dally, Petr Slovak, Julian Edbrooke-Childs, Max Van Kleek, and Nigel Shadbolt. I make up a silly name' understanding children's perception of privacy risks online. In *The 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

A Participant demographics

Table 3: Participant demographic details

Participant ID	Gender	Age	Occupation	Highest/Current level of education	Most recent program of study	Cyber challenges
P1-S3	Female	18-24	Student	Undergraduate	Prefer not to answer	No exposure
P2-S3	Female	35-44	Project Manager	Graduate degree	Project Management / Health Informatics	No exposure
P3-S3	Male	18-24	Student	Undergraduate	Computer Science	Completed one or more
P4-S3	Female	18-24	Student	Undergraduate	Computer Science	Completed one or more
P5-S5	Female	25-34	Student, web developer	Graduate degree	Computer Science	Completed one or more
P6-S3	Female	55-64	IT support specialist	Graduate degree	Control Systems Engineering	No exposure
P7-S2	Male	55-64	Contract Instructor	Graduate degree	No answer	No exposure
P9-S3	Male	18-24	Student	Undergraduate	Software Engineering	No exposure
P10-S3	Male	No answer	IT support specialist	Undergraduate	Engineering	No exposure
P11-S2	Female	25-34	Project Manager	Undergraduate	Business	No exposure
P12-S2	Male	65-74	Retired	Graduate degree	Applied Science (computer related)	No exposure
P13-S3	Male	25-34	Network Maintenance	Undergraduate	Computer Engineering	No exposure
P14-S3	Male	35-44	Software Engineer	Undergraduate	Computer Engineering	No exposure
P15-S3	Female	35-44	Software Developer	Undergraduate	Computer Science	No exposure
P16-S3	Male	18-24	Student	Undergraduate	Computer Science	No exposure
P18-S2	Female	18-24	Prefer not to answer	Undergraduate	Computer Engineering	No exposure
P19-S4	Male	25-34	IT support specialist	College	Information System Support Specialists	No exposure
P20-S3	Male	18-24	Student	Undergraduate	Computer Science	No exposure
P22-S3	Female	25-34	UX Designer	Graduate degree	Human Computer Interaction	No exposure
P23-S2	Male	18-24	Student	Undergraduate	Computer Systems Engineering	No exposure
P24-S2	Male	18-24	Student	Undergraduate	Computer Science (Minor: Entrepeneurship)	Completed one or more
P25-S2	Female	25-34	Student	Undergraduate	Computer Science	Completed one or more
P26-S2	Female	18-24	Student	Undergraduate	Computer Science (Stream: Software Engineering)	No exposure

B Data analysis timeline

Table 4: A summary of the data analysis process, and the researchers involved at each stage.

			Researchers involved		
Activity	**Codebook used**	**Transcripts coded**	**RS1**	**RS2**	**RS3**
Formed initial codebook of 9 items while editing all transcripts for accuracy	C1	–	x		
Coded two transcripts **(refined codebook)**	C1	T1-T2	x		
Re-coded two transcripts	C2	T1-T2	x		
Coded five transcripts	C2	T4, T6-T7, T10-T11	x		
Coded three transcripts **(refined codebook)**	C2	T3, T5, T9	x	x	
Re-coded transcripts	C3	T3, T5, T9	x	x	
Coded three transcripts **(codebook finalised)**	C3	T13, T15, T22	x	x	
Re-coded three transcripts (no changes)	C4	T3, T9, T15	x		x
Re-coded ten transcripts	C4	T1-T2, T4-T7, T10-T11, T13, T22			x
Coded seven new transcripts	C4	T12, T14, T16, T18-T20, T22-T23			x
Coded three new transcripts	C4	T24-T26	x		

C Post-test questionnaire

Q1 Please enter your Participant ID : _____

Q2 What gender do you most closely identify with?

- ○ Male (1)
- ○ Female (2)
- ○ Other: (3) _____
- ○ Prefer not to say (4)

Q3 What is your age? If you prefer not to say, please enter "prefer not to say": _____

Q4 Choose either the level of education for which you are **currently enrolled** or the highest level of education you have completed.

- ○ Elementary school (1)
- ○ High school (2)
- ○ College (3)
- ○ Technical, trade school, vocational training, or apprenticeship (4)
- ○ Undergraduate degree (Bachelor's) (5)
- ○ Post-graduate certificate or diploma (6)
- ○ Graduate degree or professional degree (7)
- ○ Other (8): _____
- ○ Prefer not to say (9)

Q5 What is your occupation?
If you prefer not to say, please enter "prefer not to say": _____

Q6 Please list all current and previously completed programs of study.
If you prefer not to answer, please write in "prefer not to answer": _____

Q7* Have you ever taken any technical **courses** or **training**? This would include courses from Computer

* Question not present in the original study's methodology.

Science, Information Technology (IT), Software Engineering, Systems Engineering, and many other fields.
Please list any other formal **training**, **courses**, or otherwise that may count as "technical".
If you prefer not to answer, please write in "prefer not to answer":

_____

Q8 Which option best describes your current employment status?

- ○ Working for payment or profit (1)
- ○ Unemployed (2)
- ○ Home-maker (looking after home/family) (3)
- ○ Student (no other form of employment) (4)
- ○ Retired (5)
- ○ Unable to work due to permanent sickness/disability (6)
- ○ Other (specify): (7) _____
- ○ Prefer not to say (8)

Q9* Do you study or work in a field that links closely to some form of computer **security**? (E,g: involving encryption, hacking, authentication)

- ○ Study only (please provide details into your program/area of study): _____
- ○ Work only (please provide details into your area of work): _____
- ○ Both study and work (please provide details into your area of study and work:) _____ (3)
- ○ Neither study nor work (4)
- ○ Prefer not to say (5)

Q10* Have you ever participated in hackathons or other security-oriented coding challenges?

- ○ Yes (please list what kinds of challenges you've participated in:) _____ (1)
- ○ No (2)
- ○ Prefer not to say (3)

Q11* Please list any courses you have taken pertaining to computer security.
If you prefer not to answer, please write "Prefer not to answer": _____

* Question not present in the original study's methodology.

Q12* Are you currently a student (part-time or full-time)

- ○ Yes (1)
- ○ No (2)
- ○ Prefer not to say (3)

Display The Question Below: If "Are you currently a student??" is "Yes"

Q13 Please select the level of education you are **currently completing**.

- ○ Undergraduate degree (1)
- ○ Master's degree (2)
- ○ PhD degree (3)
- ○ Post-doc (4)
- ○ Diploma (5)
- ○ Other (please list): _____ (6)
- ○ Prefer not to say (7)

Display The Question Below: If "Are you currently a student?" is "Yes"

Q14* Please enter which year of study you are currently in (e.g: 1st year, 2nd year, etc).
If you prefer not to answer, please write "Prefer not to answer": _____

Q15* Please select the statement that best describes your **exposure** to topics of computer security (encryption, hacking, authentication, etc) **in the past one year**.

- ○ No exposure at all (1)
- ○ A little exposure (2)
- ○ Some exposure (3)
- ○ High exposure (4)
- ○ Very high exposure (5)
- ○ Prefer not to say (6)

* Question not present in the original study's methodology.

Q16* Please select the statement that best describes your **level of knowledge** of computer security (encryption, hacking, authentication, etc).

- ○ No knowledge at all (1)
- ○ A little bit of knowledge (2)
- ○ Some knowledge (3)
- ○ High level of knowledge (4)
- ○ Very high level of knowledge (5)
- ○ Prefer not to say (6)

Q17 Please enter the number of hours you typically spend on each of the following activities in the specified time range.
If you prefer not to say, please enter the letter **X**.

- Recreational TV: ____ hours/week (1)
- Newspapers: ____ hours/week (2)
- Podcasts: ____ hours/week (3)
- Social media: ____ hours/**day** (4)
- Movies: ____ hours/**month** (5)
- TV news: ____ hours/week (6)
- Magazines: ____ hours/week (7)

Q18 Please select which of the following genres you enjoy consuming media in (select as many as apply).

- ☐ Action (1)
- ☐ Comedy (2)
- ☐ Romance (3)
- ☐ Documentary (4)
- ☐ Horror (5)
- ☐ Drama (6)
- ☐ Kids (7)
- ☐ Adventure (8)
- ☐ Sci-fi (9)
- ☐ Fantasy (10)
- ☐ Thrillers (11)
- ☐ Spy-films (12)
- ☐ Other (please list): _____ (13)
- ☐ Prefer not to say (14)

* Question not present in the original study's methodology.

Comparing Security and Privacy Attitudes Among U.S. Users of Different Smartphone and Smart-Speaker Platforms

Desiree Abrokwa, Shruti Das, Omer Akgul, and Michelle L. Mazurek
University of Maryland

Abstract

Many studies of mobile security and privacy are, for simplicity, limited to either only Android users or only iOS users. However, it is not clear whether there are systematic differences in the privacy and security knowledge or preferences of users who select these two platforms. Understanding these differences could provide important context about the generalizability of research results. This paper reports on a survey (n=493) with a demographically diverse sample of U.S. Android and iOS users. We compare users of these platforms using validated privacy and security scales (IUIPC-8 and SA-6) as well as previously deployed attitudinal and knowledge questions from the Pew Research Center. As a secondary analysis, we also investigate potential differences among users of different smart-speaker platforms, including Amazon Echo and Google Home. We find no significant differences in privacy attitudes of different platform users, but we do find that Android users have more technology knowledge than iOS users. In addition, we find evidence (via comparison with Pew data) that Prolific participants have more technology knowledge than the general U.S. population.

1 Introduction

The increasing ubiquity of mobile and IoT devices has generated significant research and development related to privacy and security tools, affordances, and preferences. For example, researchers have explored, at length, the implication of built-in permissions systems that govern mobile apps' access to location, contacts, sensors like the microphone or camera, and other potentially sensitive resources (e.g., [6, 9, 17, 33, 34, 38, 53, 64]). Much time and effort have also been spent developing and testing different smartphone authentication mechanisms (e.g., [31, 37, 43, 49, 60]). Extensive research into modern secure communication has focused on mobile messenger apps, including for example exploration of the usability of authentication ceremonies [24, 28, 41, 57, 61, 62, 66].

In the IoT ecosystem, researchers have explored issues ranging from concerns about unexpected listening and recording [32, 36, 55] to attacks requiring user interaction [29, 50], to studies of IoT privacy and security concerns more generally (e.g., [3, 15, 56, 67]), and more.

In many cases, these studies have been limited — often for simplicity or convenience — to only one mobile or IoT platform (e.g., Android or the Amazon Echo ecosystem) [5, 9, 17, 29, 34, 54, 59, 61, 64, 66]. In other cases, researchers have supported multiple platforms, at the cost of more complicated study instruments that must work in multiple settings [4, 36, 50, 62].

Given this context, it it important to know whether there are meaningful differences in privacy and security preferences, beliefs, and attitudes between users of different platforms. For example, Apple has recently marketed its products as more privacy-protective than alternatives [2]. In the past, iOS has pioneered fine-grained permission controls, including limiting location permissions to single-use or only while an app is being used [63]. In contrast, Google's largest source of income[1] is though targeted advertising, involving extensive user data collection.

We hypothesize that this distinction in business strategies could result in more privacy-sensitive consumers tending to purchase iPhones, perhaps resulting in Android users who are disproportionately unconcerned with privacy. Similar questions are also applicable to smart speaker platforms; however, market positioning related to privacy is not (yet) as clear as with smartphones. If there are indeed meaningful differences between users of different platforms, then extra work by re-

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

[1] https://abc.xyz/investor/static/pdf/2020Q4_alphabet_earnings_release.pdf

searchers to ensure their studies support multiple platforms may be critical. On the other hand, if there are not meaningful differences, then researchers can opt for simpler experimental designs with less concern about reduced generalizability.

In this paper, as our primary objective, we address these questions by surveying privacy and security attitudes among users of different mobile and IoT platforms to determine if differences exist. We use validated scales to measure security attitudes (SA-6) and privacy concern(IUIPC-8) [16, 19]. We also reuse questions previously used by the Pew Research Center (henceforth: Pew) in a nationally representative survey to ask about skepticism toward company data practices and knowledge about digital privacy and security [1]. As a secondary objective, reusing these questions allows us to compare attitudes between Prolific and nationally representative samples.

In 2014, Reinfelder et al. addressed similar questions, comparing security behaviors between Android and iOS users in Germany, finding that Android users were somewhat more privacy- and security-conscious [48]. We revisit this question to see what has changed in the intervening years, as devices have evolved and Apple has marketed privacy more heavily. In addition to smartphone platforms, we also consider the increasingly important smart-speaker platform. Further, we deliberately focus on attitudes rather than behaviors, as we expect that behaviors are more likely to be influenced by different platforms' privacy and security affordances.

To ensure a diverse sample, we recruit 493 participants using Prolific's "representative" sample feature, which approximates the U.S. population for gender, race, and age. We find no significant differences in security attitudes, privacy concern, or skepticism toward company data practices between users of different mobile or IoT platforms. We do find that Android users score slightly higher in security and privacy knowledge than iOS users. We also compare our sample to the representative Pew sample for the two Pew metrics, finding no difference in skepticism; however, our participants scored significantly higher in security and privacy knowledge, somewhat limiting the generalizability of our primary analysis.

These findings have implications for the design of future research exploring uses and preferences in mobile and IoT security and privacy. For studies purely about attitudes and preferences, ensuring cross-platform representation may not be necessary. On the other hand, for studies where knowledge may play an important role — for example, in evaluating mental models of security and privacy mechanisms — ensuring participation from both iOS and Android users may be more important.

2 Related Work

We discuss related work in two key areas: metrics for privacy and security, and studies that compare privacy or security attitudes and preferences among various populations.

Privacy and security metrics Researchers have long sought to define metrics for privacy and security attitudes as well as behavior. Several developed psychometric scales intended to measure privacy attitudes and concern. Perhaps the first such scale was the original 1991 Westin Privacy Segmentation Index, which groups respondents into privacy fundamentalists, pragmatists, and unconcerned [30]. In 1996, Smith et al. developed the Concern for Information Privacy (CFIP) scale, which measured privacy concern along multiple dimensions including collection, unauthorized secondary use, and improper access [52]. This was followed in 2004 by the IUIPC, a 10-item scale that builds on the original CFIP and measures three dimensions of privacy attitudes: control, awareness of privacy practices, and collection [35]. A number of other privacy scales have been proposed; Preibusch provides a comprehensive list and comparison [42].

Other researchers have investigated the utility and reliability of these scales. Woodruff et al. demonstrated that the Westin index is poorly predictive of privacy-relevant behavioral intentions [65]. In 2013, Preibusch's aforementioned guide reviews pros and cons of each metric before finally recommending IUIPC [42]. However, Sipior et al. and Zeng et al. obtain mixed results when re-validating the IUIPC, particularly with respect to trust in online companies and social networking, respectively [51, 68].

Most recently, Groß demonstrated that the original IUIPC-10 contains two poorly worded questions, without which the scale is significantly more reliable [19]. In this work, we adopt the resulting IUIPC-8 scale.

Other scales concern security attitudes and behaviors. The Security Behavior Intentions scale (SeBIS) by Egelman and Peer is intended to measure how well individuals comply with computer security advice from experts [14]. Faklaris et al. created and validated the six-item SA-6 scale to measure security attitudes, which may differ from (intended) behaviors [16]. Because we focus primarily on attitudes, we select SA-6 rather than SeBIS for our study.

Security- and privacy-relevant questions also appear in regularly administered, representative-sample surveys conducted by the Pew Research Center. The center's 2019 American Trends Panel: Wave 49 features relevant questions related to Americans' knowledge of web and internet concepts, as well as questions related to skepticism (or trust) that companies will manage the data they collect appropriately [1]. We adopt subsets of these questions that align with our research goals. Including these questions allows us to compare our results to a fully representative random sample of U.S. adults.

Comparing sample populations for privacy and security Other research has sought to compare privacy and security attitudes among different populations. Kang et al. compared the privacy attitudes and behaviors of U.S. Mechanical Turk (MTurk) workers with the general U.S.

population, finding U.S. MTurk workers display heightened privacy attitudes [25]. Redmiles et al. endorse the use of MTurk workers for convenient, affordable samples. However, they highlight shortcomings when trying to generalize security and privacy perceptions of underrepresented groups (e.g. elderly, less educated) [46]. Because we employ questions from Pew, we are able to similarly compare our results to the broader U.S. population [1].

Research has also explored differences in privacy attitudes and preferences in different countries and regions. In a longitudinal study that included 25 countries, Kelley identified important regional differences in the importance people assign to privacy, as well as whether and when it is acceptable for, e.g., law enforcement organizations to violate privacy in pursuit of other goals [26]. Redmiles compared behavior after Facebook security incidents in five countries, finding some cultural differences [45]. Ion et al. noticed political and cultural attitudinal differences in mental models related to cloud computing privacy and security between Swiss and Indian communities [23]. Similarly, Harbach et al. studied more than 8,000 Android users across eight countries. Their results affirmed that cultural and demographic characteristics can strongly determine security and privacy considerations [21]. Dev et al. compared privacy concerns related to Whatsapp messaging in Saudi Arabian and Indian communities, finding likely culturally influenced behavioral differences between populations but overall similar privacy trends when considering participants within each sample [12].

Most closely related to our work are three separate 2013-2014 studies comparing security and privacy awareness between Android and iOS users. In the first, King interviewed a small sample of iPhone and Android users from San Francisco to qualitatively understand contextual design decisions that impact privacy-centered user experiences [27]. In the second, Reinfelder et al. found (among German university students) Android users were more likely to be security aware and privacy conscious [48]. Finally, Mylonas et al. investigated user mental models of application installations on different platforms among Greeks [39]. Although not the primary research objective, they provided evidence that Android users were more security aware across multiple metrics (e.g., likelihood of adopting security software).

Because of the rapid changes in smartphone technology, both hardware and software, over the last seven years, we wanted to evaluate whether these results would still hold, this time across a broad U.S. sample. Both King and Reinfelder et al. focused on behavioral patterns, such as installing security updates, consciousness of possible malware infections, and app permissions [27,48]. We instead focus on attitudinal questions, which are frequently used in studies of smartphone and IoT users [8,10,15,44]. Further, behavioral questions about, e.g., app permissions are difficult to entangle from system design affordances and nudges that may contribute to users of different platforms making different choices. Additionally, Reinfelder et al. and Mylonas et al. primarily sampled young people. In contrast, we use Prolific's "representative sample" feature to obtain participants of diverse ages across a quasi-representative U.S. sample [39,48].

Other fields have also compared Android and iOS users. Psychologists found socioeconomic factors and personality traits may contribute to smartphone preferences [20].

3 Methods

To answer our research questions, we created and distributed a survey to measure the privacy and security attitudes and perceptions of participants. The survey was approved by the University of Maryland's Institutional Review Board. Our experimental approach was also preregistered with AsPredicted. [2]

In the following subsections we discuss the survey design, our recruitment process, our data analysis approach, and the limitations of our study.

3.1 Survey

We designed a short survey measuring privacy and security attitudes and perceptions, building on various previously used and validated constructs as described in Section 2.

The survey included the SA-6 [16] and the IUIPC-8 [19], as well as four questions about skepticism toward data use by companies and seven security- and privacy-relevant knowledge questions, all taken from Pew [1]. The original Pew survey contained 10 digital knowledge questions; we used seven that are privacy- and security-relevant. For example, we selected questioned related to HTTPS, private browsing, and phishing, while deeming a question asking participants to identify a technology leader from their photo irrelevant. To distinguish the two sets of Pew questions, we refer to them going forward as the *skepticism* and *knowledge* metrics, respectively. The questions chosen for the skepticism and knowledge metrics are shown in Tables 1 and 2.

To ensure that the Pew skepticism questions could be added together for use as a single consistent metric, we tested their internal reliability with Cronbach's α, using the data collected in Pew's national survey. We obtained $\alpha = 0.83$ for the four skepticism questions: above the 0.80 threshold for "good" reliability [18].

After providing consent, participants provided their country of residence, as a confirmation of Prolific's selection criteria. As we intended to recruit only U.S. participants, those who answered with other countries were filtered out immediately.

Next, we asked for background information on participants' device(s) and how they use them. This included multiple-choice questions about how many smartphones the participant uses or owns, what purposes they use their smartphone for

[2] https://aspredicted.org/gx2v9.pdf

Item ID	Item Text
PP5A	Follow what their privacy policies say they will do with your personal information
PP5B	Promptly notify you if your personal data has been misused or compromised
PP5C	Publicly admit mistakes and take responsibility when they misuse or compromise their users' personal data
PP5D	Use your personal information in ways you will feel comfortable with

Table 1: Items related to skepticism of company data practices, drawn from the Pew Research Center American Trends Panel: Wave 49 questionnaire [1], that are included in our survey. In the survey, participants are asked: How confident are you, if at all, that companies will do the following things? Response options are a four-point Likert-type scale from very confident to not confident at all. The item IDs are those used by Pew.

(e.g., personal, work, other), the model and operating system of their primary smartphone, whether or not they own a smart speaker (and if so, which one), how frequently they use the voice assistant on their smartphone, and how frequently they use their devices (e.g., multiple times a day). Participants were asked to retrieve actual time-use data from their smartphone if applicable. Participants without a smartphone were filtered out at this point.

Next, participants answered the security and privacy perceptions questions, including SA-6, IUIPC-8, the Pew skepticism metric and the Pew knowledge metric. In keeping with their original use, we randomized the question order and answer choices within the Pew segments. We also randomized the order of the three IUIPC-8 subscales (but not the order of questions or answers within subscales). This section also included free-response questions asking participants to explain their choices for two questions; these responses were used primarily as attention checks, and participants who gave unrelated or non-responsive answers to these questions were removed from the sample.

Finally, we asked some standard demographic questions, including questions related to age, gender identity, race/ethnicity, and employment status. We also asked about tech-savviness, measured using a Likert-type question about how often the participant gives technology advice to others. The full survey text is given in Appendix B.

We implemented the survey in Qualtrics. Prior to main data collection, we conducted eleven pilot tests of the survey with a convenience sample, to validate the questions and survey flow, as well as to estimate the time required for completion (15 minutes).

3.2 Recruitment

Participants were recruited through Prolific, an online crowdsourcing platform which can be expected to produce high-quality results [40]. Participants were required to reside in the United States and be 18 or older. The study was advertised as being about "Technology Perceptions" to avoid self-selection biases related to privacy and/or attachment to different hardware vendors. We used Prolific's "representative sample" tool to increase the diversity of our sample. Prolific stratified our sample to match 95% of 2015 U.S. census values for age, gender, and race [58].

Participants who completed the survey with valid responses were compensated with $3.00. The survey took on average 12.4 minutes, resulting in average compensation of of $14.56/hour. Responses were collected in December 2020.

3.3 Analysis

We analyzed our data using four linear regression models, with dependent variables for each privacy/security metric: SA-6, IUIPC-8, the Pew skepticism metric, and the Pew knowledge metric. For SA-6, IUIPC-8, and the skepticism metric, we summed participants' Likert responses. For the knowledge metric, participants were scored 0 to 7 based on how many questions they answered correctly.

For all four models, the independent variables included smartphone platform (iOS or Android) and smart-speaker platform (Amazon Echo, Google Home, other, none). Other covariates included age, gender, daily estimated smartphone use time, whether or not the smartphone was rooted/jailbroken, and how often participants give tech advice (used as a proxy for tech savviness). For parsimony, we binned tech advice responses into two categories: *less often* (never rarely, sometimes) and *more often* (often, almost always). We similarly binned gender into men and *non-men* (women and other genders), because very few participants reported other genders. These variables are summarized in Table 3.

To obtain our four models, we perform model selection based on Akaike Information Criterion (AIC), which strikes a balance between how well models explain the dependent variables and over-fitting [7]. For each dependent variable, we fit regressions with smartphone and smart-speaker platforms (the main variables of interest) as well as all possible combinations of the other covariates. We report only the model with the lowest AIC for each metric.

We aimed to recruit 500 participants. Power analysis for linear regression (assuming that all our potential IVs would be included) shows that 500 participants is sufficient to detect approximately small[3] effects ($f^2 = 0.032$) [11].

We note one deviation from our preregistered analysis plan. We initially planned to fit eight regression models: one for

[3] Cohen claims $f^2 > 0.02$ would capture "small", $f^2 > 0.15$ would capture "medium" effect sizes.

Item ID	Item text	Correct answer
KNOW1	If a website uses cookies, it means that the site …	Can track your visits and activity on the site
KNOW2	Which of the following is the largest source of revenue for most major social media platforms?	Allowing companies to purchase advertisements on their platforms
KNOW3	When a website has a privacy policy, it means that the site …	Has created a contract between itself and its users about how it will use their data
KNOW4	What does it mean when a website has "https://" at the beginning of its URL, as opposed to "http://" without the "s"?	Information entered into the site is encrypted
KNOW5	Where might someone encounter a phishing scam?	All of the above (In an email, on social media, in a text message, on a website)
KNOW7	The term "net neutrality" describes the principle that …	Internet service providers should treat all traffic on their networks equally
KNOW8	Many web browsers offer a feature known as "private browsing" or "incognito mode." If someone opens a webpage on their computer at work using incognito mode, which of the following groups will NOT be able to see their online activities?	A coworker who uses the same computer

Table 2: Security- and privacy-relevant digital knowledge questions, drawn from the Pew Research Center American Trends Panel: Wave 49 questionnaire [1]. All questions are multiple-choice. The item IDs are those used by Pew.

Variable	Explanation	Baseline
Main variables of interest:		
Smartphone OS	Whether the participant is an iOS or Android user	iOS
Smart speaker	Whether the participant owns a smart speaker, and which	Amazon Echo device
Demographic covariates:		
Tech advice	Whether the participant is asked for tech advice, binned into less often or more often	Less often
Device rootedness	Whether or not the participant's device is rooted or jailbroken	Not rooted
Screen-time estimate	Self-reported hours of daily phone use	–
Age	The participant's age	–
Gender	Gender, binned into men and non-men (women and other genders)	Non-man

Table 3: Independent variables (**IVs**) used in our regressions, including main variables of interest (mobile and smart-speaker platforms) as well as demographic covariates. Baselines are listed for categorical variables. Section 3.3 details the regressions.

each combination of dependent variable (SA-6, IUIPC, skepticism metric, knowledge metric) and platform (smartphone OS and smart-speaker type). We made this plan because we assumed that relatively few participants in the initial "representative sample" from Prolific would own smart speakers; we intended to augment our sample with a second batch recruited from Prolific specifically on the basis of smart-speaker ownership. Because we didn't want to combine these two incompatible samples, we intended to model smartphone and smart-speaker platforms separately. However, we were pleasantly surprised to find that more than one third of our "representative" sample were smart-speaker owners. Rather than obtain a less representative sample, we opted to use only the initial sample and to include both platform types in our four regression models. Using fewer models reduces the complexity of our analysis and enables holistic comparison that accounts for all factors at once.

We also added one secondary analysis not described in our pre-registration: We compare our participants' responses to the Pew questions to the nationally representative Pew data for the same questions. This comparison allows us to explore how well the "representative" Prolific feature captured the broader U.S. population (albeit with a time lag). For these comparisons, after establishing that the data was not normally distributed (Shapiro-Wilk $p < 0.001$), we use non-parametric, two-tailed Mann-Whitney U tests, one for each Pew metric. Since the Pew scales are used in two analyses each (one regression and one MWU), we adjust the relevant p-values with Bonferroni correction.

3.4 Limitations

Our study has several limitations, most of which are common to this type of research. Although we used Prolific's "represen-

tative sample" [4] tool to diversify our sample, our participants are still on average more educated than the U.S. population. Our sample also severely underrepresents, compared to the U.S. population, people who identify as Hispanic or Latino; the Prolific stratification does not incorporate this ethnicity information. Additionally, we compared the results for the Pew scales to a representative sample of the U.S. population. This indicated that while Prolific users have similar privacy concerns, they have more privacy and security knowledge than the broader population.

Survey responses were only collected from Prolific users in the United States. We focused on the United States to avoid confounds related to availability and popularity of different devices, as well as cultural differences, inherent in comparing multiple countries. However, our results cannot necessarily generalize to non-U.S. populations.

We use self-report metrics, which are vulnerable to biases such as social desirability and acquiescence. However, prior work suggests self-reporting can provide useful data on security and privacy questions [13,47]. Further, we expect these biases to affect users of different smartphone and smart-speaker platforms similarly, enabling comparison among groups.

4 Results

In this section, we first describe our survey participants. We then detail the results of our regressions comparing platform users across each security or privacy metric. Finally, we compare our sample to the nationally representative Pew sample for context. Overall, we found no differences across platforms in privacy attitudes, but we found that Android users scored higher than iOS users on the Pew knowledge metrics. Our sample did not differ significantly from the Pew sample in skepticism toward company data practices, but our participants scored higher on the knowledge metric.

4.1 Participants

In December 2020, we used Prolific to recruit 500 participants currently residing in the U.S. We discarded five for off-topic or unresponsive free-text responses, one for being outside the U.S., and one who skipped an optional question that was required for our analysis. The remaining 493 participants served as our final sample for analysis.

Of the 493 participants, 285 use Android and 208 use iOS on their primary smartphone. In total, 175 participants use smart speakers, including 95 who only use an Amazon Echo, 54 who only use Google Home, and 26 who use some other smart speaker or use multiple brands.

Demographics within our Android, iOS, and total samples are given in Table 4. Because we used Prolific's "representative sample" feature, our overall sample is fairly representative of the U.S. for gender, age, and race. Other demographics, however, suffer from typical crowdsourcing biases, including insufficient Hispanic/Latinx representation and more education than the U.S. population overall[5]. The plurality of participants report giving tech advice "sometimes."

		Android (%)	iOS (%)	Total (%)
Gender	Women	50.5	51.0	50.7
	Men	48.4	47.6	48.1
	Non-binary and other	1.1	1.5	1.2
Age	18-27	14.0	24.0	18.3
	28-37	19.7	16.8	18.5
	38-47	18.2	14.9	16.8
	48-57	17.2	18.3	17.6
	58+	30.9	26.0	28.8
Hispanic origin	No	95.8	91.8	94.1
	Yes	4.2	8.2	5.9
Race	White	74.0	73.3	73.7
	Black or African Amer.	15.5	11.5	13.8
	Asian	6.1	11.1	8.2
	Amer. Ind. or AK Native	2.7	1.8	2.3
	Nat. Hawaiian or Pac. Isl.	0.3	0.5	0.4
Education	Completed H.S. or below	13.0	5.3	9.7
	Some college, no degree	25.3	22.6	24.1
	Associate's degree	12.6	4.8	9.3
	Bachelor's degree	27.7	36.1	31.2
	Master's degree or higher	14.4	25.0	18.9
Employment status	Employed full-time	34.4	37.5	35.6
	Employed part-time	13.0	13.5	13.2
	Self-employed	13.0	12.5	12.8
	Retired	15.8	13.9	15.0
	Unemployed	7.7	6.7	7.3
	Student	5.6	9.6	7.3
Tech advice	Almost always	5.6	6.7	6.1
	Often	20.3	19.2	19.9
	Sometimes	41.4	43.3	42.2
	Rarely	28.1	24.5	26.6
	Never	4.6	6.3	5.3

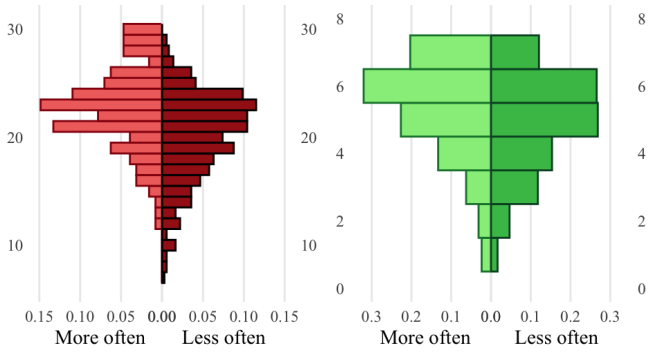
Table 4: Participant demographics. Percentages may not add to 100% due to multiple selection and item non-response; some categories with small percentages are elided.

Our results show some demographic differences between smartphone users. Our Android users tend to be older and less educated than their iOS counterparts. Our iOS sample has higher proportions of Asian and Hispanic people, but a noticeably smaller proportion of Black people. There are also some notable differences in educational attainment between the populations, with iOS users tending to have more education.

In addition to asking participants for their daily screen-time estimates, we asked participants who were able to report their actual daily screen-time averages (visible under "Screen Time" settings on iOS and "Digital Wellbeing" on some An-

[4] https://researcher-help.prolific.co/hc/en-gb/articles/360019236753-Representative-Samples-on-Prolific

[5] https://www.census.gov/acs/www/data/data-tables-and-tools/data-profiles/2019



(a) SA-6 (b) Pew knowledge

Figure 1: Comparison of SA-6 and Pew knowledge metric responses for participants who give tech advice more and less often. The Y-axis shows the range of possible values for each scale; the X-axis shows the fraction of total participants with each score. Both comparisons show a significant difference.

SA-6	β	$CI_{95\%}$	T-value	p-value
Smartphone (vs. iOS)				
Android	0.0	[-0.7, 0.8]	0.070	0.945
Smart speaker (vs. Amazon)				
Google	−0.4	[-1.8, 1.0]	−0.600	0.549
Other	−0.9	[-2.7, 0.8]	−1.026	0.306
None	−0.1	[-1.0, 0.9]	−0.196	0.845
Covariates				
Tech advice: More often	2.5	[1.7, 3.4]	5.901	< 0.001*
Rootedness: Not rooted	−1.1	[-2.6, 0.4]	−1.482	0.139
Gender: Man	−0.9	[-1.7,-0.2]	−2.533	0.012*

Table 5: Final regression table for SA-6. Adj. $R^2 = 0.08$. *Statistically significant.

droid models). Self-reported daily screen time was 4.5 hours ($\sigma = 3.7$, min=0, max=20). Participants on average (calculated from 225 participants who were able to provide both an estimate and the smartphone report) underestimated their screen time by 27.4 minutes ($\sigma = 147.3$). This corresponds to 10% error in screen-time use. Distribution of the error can be found in Figure 4 of Appendix A.

4.2 Comparing platforms

We fit four regression models, one each for our privacy and security metrics. These models included both smartphone and smart-speaker platform, as well as other demographic covariates. We report, in turn, on each of the final best-fit models.

Security attitude (SA-6) First, we analyze responses to the SA-6 security attitude scale. Potential scores on this scale range from 6–30, with higher numbers indicating a more positive attitude toward security behaviors. Overall, our participants scored an average of 20.7 ($\sigma = 4.2$, min=7, max=30).

By definition, our final regression model (Table 5) includes both smartphone (Android mean=20.8; iOS mean=20.7) and smart-speaker platform (Amazon Echo mean=20.9; Google Home mean=20.8; Other mean=20.4; None mean=20.7), but neither factor is significant. Figure 2a illustrates the similarity between iOS and Android participants for this metric.

The only two significant covariates were tech advice and gender. As shown in Figure 1a, those who give tech advice "often" or "almost always" were associated with a 2.5-point increase in positive attitude ($p < 0.001$), compared to those who do not. It's intuitively reasonable that increased tech-savviness would correlate with more interest in security. This also aligns with findings in the original SA-6 paper that the scale correlates with tech-savviness, confidence in using computers, and digital literacy [16].

In a smaller effect, men were associated with an 0.9-point decrease in positive attitude toward security compared to non-men ($p = 0.012$). Rootedness was also retained in the final model, but did not show a statistically significant effect.

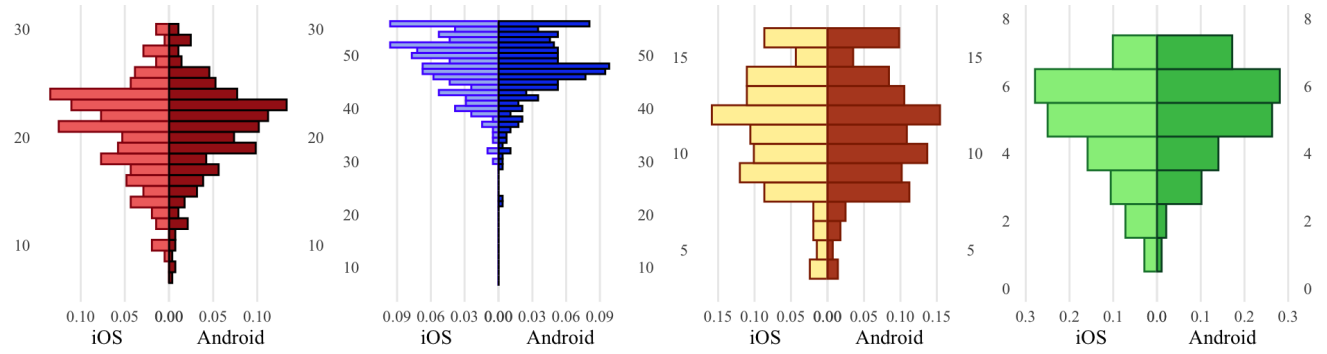
IUIPC-8	β	$CI_{95\%}$	T-value	p-value
Smartphone (vs. iOS)				
Android	−0.9	[-1.9, 0.2]	−1.624	0.105
Smart speaker (vs. Amazon)				
Google	−1.4	[-3.4, 0.5]	−1.418	0.157
Other	−0.4	[-2.9, 2.2]	−0.274	0.784
None	1.0	[-0.3, 2.4]	1.517	0.130
Covariates				
Rootedness: Not Rooted	1.6	[-0.6, 3.7]	1.449	0.148
Screen-time Estimate	−0.2	[-0.3, 0.0]	−2.223	0.027*
Age	0.0	[-0.0, 0.1]	1.510	0.132

Table 6: Final regression table for IUIPC. Adj. $R^2 = 0.04$. *Statistically significant.

Privacy concern (IUIPC-8) Next, we consider responses to the IUIPC-8, which measures privacy concern. Potential scores range from 8–56, with higher scores indicating higher levels of privacy concern. Our participants scored on average 47.7 ($\sigma = 5.9$, min=22, max=56), indicating that they tend to be more privacy sensitive than not.

As with SA-6, we see no significant differences based on smartphone (Android mean=47.3; iOS mean=48.1) or smart-speaker platform (Amazon Echo mean=47.2; Google Home mean=45.3; other mean=46.8; None mean=48.3). Figure 2b illustrates the similarity of responses across the two smartphone platforms and Table 6 shows the final regression model.

In fact, we find only one significant factor: estimated screen time on the primary smartphone, depicted in Figure 5a of Appendix A. Each additional 5 hours of daily screen time is associated with a drop of 1.0 points in privacy concern



(a) SA-6 (b) IUIPC-8 (c) Pew skepticism (d) Pew knowledge

Figure 2: Comparison between iOS and Android users on all metrics. The Y-axis shows the range of possible values for each scale; the X-axis shows the fraction of total participants with each score. Only the Pew knowledge metric shows a significant difference between platforms.

($p = 0.027$). It is perhaps unsurprising that participants who spend more time on their smartphones exhibit lower privacy concern, possibly due to habituation. Age and whether or not the participant had rooted their device were retained in the final model but did not show significant effects.

Skepticism toward companies (Pew) We next examine participants' skepticism toward companies' data management practices. Potential scores on this metric range from 4–16, with higher scores indicating more skepticism and lower scores indicating more trust. Participants scored on average 11.3 ($\sigma = 2.8$, min=4, max=16).

On this metric, also, we see no significant differences based on smartphone (Android mean=11.2; iOS mean=11.3) or smart-speaker (Amazon Echo mean=11.5; Google Home mean=11.1; Other mean=10.8; None mean=11.3) platform in the final model (Table 7). This lack of difference is illustrated in Figure 2c.

As with IUIPC, the only significant factor in the model was screen time. Each additional five hours of screen time per day is associated with a 1.0-point drop in skepticism ($p < 0.001$). This aligns with the similar finding for IUIPC: more screen time, and presumably more habituation, is associated with less concern about data practices (Figure 5b of Appendix A). Age is again included in the final model but not significant.

Security and privacy knowledge (Pew) Finally, we analyzed results from the Pew knowledge metric. Participants could score from 0–7 on this metric, corresponding to the number of questions they answered correctly. Our participants scored on average 5.0 ($\sigma = 1.5$, min=0, max=7).

The final model (Table 8) estimates that Android users are likely to score 0.4 points higher on this correctness quiz than iOS users ($p = 0.004$), meaning they have somewhat more security and privacy knowledge (Android mean=5.1; iOS mean=4.8). This difference is fairly small, but may reflect Apple's reputation of making products that are easy to use even for people with very limited technological skills. This

Pew Skepticism	β	$CI_{95\%}$	T-value	p-value
Smartphone (vs. iOS)				
Android	0.0	[-0.4, 0.5]	0.191	1.000
Smart speaker (vs. Amazon)				
Google	−0.4	[-1.3, 0.5]	−0.884	0.754
Other	−0.6	[-1.8, 0.6]	−0.952	0.683
None	−0.3	[-1.0, 0.3]	−1.046	0.593
Covariates				
Screen-time Estimate	−0.2	[-0.3,−0.1]	−5.861	< 0.001*
Age	0.0	[-0.0, 0.0]	−1.554	0.242

Table 7: Final regression table for the Pew skepticism metric. Adj. $R^2 = 0.06$. *Statistically significant. All p-values reflect Bonferroni correction.

difference can be seen in Figure 2d, which shows more Android users at the high end and more iOS users at the low end of scores. We found no significant differences among smart-speaker owners on this metric, either (Amazon Echo mean=5.1; Google Home mean=5.2; Other mean=4.5; None mean=4.9).

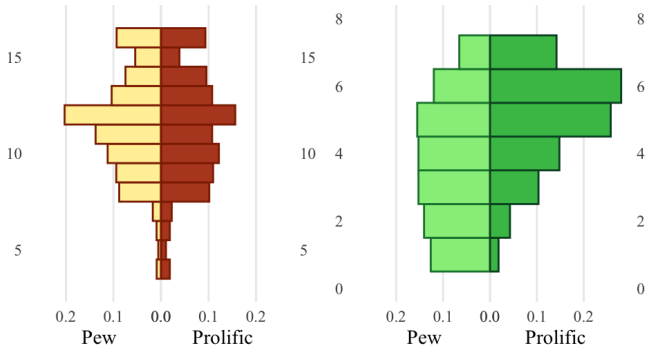
Three demographic covariates appear as significant factors in this model. Giving tech advice "often" or "almost always" (depicted in Figure 1b) correlates with an 0.4-point increase in score ($p = 0.027$); this makes intuitive sense. On the other hand, each additional five hours of screen time is associated with an 0.5-point drop in knowledge scores ($p < 0.001$). This aligns with our results on the other metrics showing that more screen time is associated with lower privacy concern and skepticism.

We also see a small but significant effect for age: each 10 years of additional age correspond to an estimated 0.1-point drop in correctness score ($p = 0.012$)[6]. It is perhaps unsur-

[6]The age coefficient (β) shown in table 7 and 8 is rounded down to 0.0;

Pew Knowledge	β	$CI_{95\%}$	T-value	p-value
Smartphone (vs. iOS)				
Android	0.4	[0.2, 0.7]	3.136	0.004*
Smart speaker (vs. Amazon)				
Google	−0.1	[-0.6, 0.4]	−0.238	1.000
Other	−0.6	[-1.3, 0.0]	−1.917	0.112
None	−0.2	[-0.5, 0.1]	−1.136	0.513
Covariates				
Tech advice: More often	0.4	[0.1, 0.7]	2.474	0.027*
Screen-time Estimate	−0.1	[-0.1, 0.0]	−4.588	< 0.001*
Age	0.0	[-0.0, 0.0]	−2.753	0.012*
Gender: Man	−0.2	[-0.5, 0.1]	−1.540	0.248

Table 8: Final regression table for the Pew knowledge metric. Adj. $R^2 = 0.08$. *Statistically significant. All p-values reflect Bonferroni correction.



(a) Pew Skepticism (b) Pew Knowledge

Figure 3: Comparison between Pew participants and our Prolific participants. The populations show significant difference in privacy/security knowledge.

prising that older people have on average slightly less security and privacy knowledge. We attribute the relatively small effect size in part to Prolific participants; in prior work, older crowdworkers and digital panel participants were unusually tech savvy for their age [46].

4.3 Comparing our participants to a nationally representative sample

An added benefit of reusing Pew questions is that we can compare responses from our sample to Pew's nationally representative sample (n=4225) [1].

Figure 3a compares our sample to the Pew sample on the skepticism metric. We find no significant difference between the two populations (MWU, $p = 0.120$).

Figure 3b illustrates responses to the knowledge metric from the two samples. Our Prolific participants tended to score higher, indicating more security and privacy knowledge (MWU, $p < 0.001$). The location-shift estimate, a measure of effect size related to median [22], is 2.0, indicates that our participants tend to score about two points higher out of seven.

5 Discussion

We used a survey with a quasi-representative sample to compare privacy and security perceptions across users of smartphone platforms (Android and iOS) as well as smart-speaker platforms (Google Home, Amazon Echo, another platform, or none). We find no significant differences in attitudes toward security, privacy, or company data practices. We do, however, find that Android users are somewhat more knowledgeable about digital security and privacy. On the other hand, differences in smartphone screen time are significantly negatively correlated with all of our metrics except security attitudes: more screen time is associated with less privacy concern, less skepticism, and less security/privacy knowledge. In a similar result, giving tech advice more often is positively correlated with positive security attitudes and more privacy/security knowledge.

These results have several implications for future research into tools and interfaces for mobile and IoT privacy and security. It may be low-effort to incorporate users of different platforms into survey or interview studies. However, cross-platform support is more challenging for research that involves new tools, such as testing an agent for managing app permissions, or field-type studies in which participants use smart devices in their homes for a period of time.

With respect to smartphones, our results suggest that studies that chiefly involve attitudes and preferences — for example, studies related to app permission choices or preferences for potentially invasive tracking and advertising — may not need to take differing platforms into account. On the other hand, we did find differences in security and privacy knowledge, which implies that cross-platform support may be important when a user's knowledge is expected to be a key factor. These could include studies evaluating knowledge or mental models related to secure communications or tracking and inferencing, as well as studies relating to adoption of various privacy- and anonymity-enhancing technologies.

Our results about screen time and tech advice also have research design implications. Many researchers already tend to (at least partially) control for tech-savviness in participants. Our results support this practice, while suggesting that screen time may be an equally or even more important variable to consider.

With respect to smart-speaker platforms, we found no significant differences in any of our metrics. This suggests that, for now, cross-platform differences are not critical for security and privacy research on smart speakers. It remains an open

when multiplied by 10, it rounds to 0.1.

question whether this result extends to other kinds of IoT devices. It is similarly unclear whether this result will remain stable over time, as the market for IoT devices becomes more mature.

Our work also has implications for crowdsourced samples. Comparing our sample to a U.S.-representative sample from Pew, we find that our participants express similar skepticism toward data practices, but are noticeably more digitally knowledgeable than the general U.S. population. The lack of difference in skepticism provides hope that the gap in privacy attitudes noted by Kang et al. in 2014 is shrinking as digital habits and devices become further entrenched [25]. On the other hand, we confirm prior results that web survey panels, even when more or less demographically representative, still provide participants who are disproportionately tech-savvy for their demographics [46]. We therefore encourage researchers to continue to recognize this limitation in generalizability, and to consider alternate means of recruiting, if feasible, when tech-savviness is important to the research question(s) being addressed.

Finally, we suggest researchers also measure other potential differences between the user populations we investigate in this study. Specifically, we emphasize the need for behavioral studies to complement our self-report data, and to explore differences between attitudes and behaviors that may relate to available privacy or security affordances.

6 Conclusion

In this study, we conducted a security and privacy survey using previously validated metrics in order to examine whether there are important differences in attitudes between users of different smartphone and smart-speaker platforms. Using a quasi-representative sample, we found no differences in attitudes among these groups. However, we found that Android users tend to have more security and privacy knowledge than iOS users. We also found that more daily screen time is associated with less privacy concern, less skepticism of company data practices, and less security and privacy knowledge. By comparing our sample to a nationally representative dataset from Pew, we can observe that our quasi-representative sample has similar skepticism to the general U.S. population, but more security and privacy knowledge. These results can provide guidance for designing — and context for interpreting — future studies on technology platforms.

7 Acknowledgments

The authors would like to thank participants who took part in our survey as well as the anonymous reviewers for constructive comments and suggestions. This paper results from the SPLICE research program, supported by a collaborative award from the National Science Foundation (NSF) SaTC Frontiers program under award number 1955805. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF. Any mention of specific companies or products does not imply any endorsement by the authors, by their employers, or by the NSF.

References

[1] American Trends Panel Wave 49, 2019. https://www.pewresearch.org/internet/dataset/american-trends-panel-wave-49.

[2] Data privacy day at Apple: Improving transparency and empowering users. 2021. https://www.apple.com/newsroom/2021/01/data-privacy-day-at-apple-improving-transparency-and-empowering-users.

[3] Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. More than smart speakers: Security and privacy perceptions of smart home personal assistants. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. USENIX Association, August 2019.

[4] Ruba Abu-Salma, Kat Krol, Simon Parkin, Victoria Koh, Kevin Kwan, Jazib Mahboob, Zahra Traboulsi, and M. Angela Sasse. The Security Blanket of the Chat World: An Analytic Evaluation and a User Study of Telegram. In *2nd European Workshop on Usable Security (EuroUSEC)*. Internet Society, NDSS Symposium, 2017.

[5] Omer Akgul, Wei Bai, Shruti Das, and Michelle L. Mazurek. Evaluating In-Workflow Messages for Improving Mental Models of End-to-End Encryption. In *USENIX Security Symposium*, 2021.

[6] Bram Bonné, Sai Teja Peddinti, Igor Bilogrevic, and Nina Taft. Exploring decision making with Android's runtime permission dialogs using in-context surveys. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*. USENIX Association, July 2017.

[7] Hamparsum Bozdogan. Model Selection and Akaike's Information Criterion (AIC): The General Theory and Its Analytical Extensions. *Psychometrika*, 1987.

[8] Alex Braunstein, Laura Granka, and Jessica Staddon. Indirect Content Privacy Surveys: Measuring Privacy Without Asking About It. *SOUPS 2011 - Proceedings of the 7th Symposium on Usable Privacy and Security*, 2011. https://dl.acm.org/doi/10.1145/2078827.2078847.

[9] Pern Hui Chia, Yusuke Yamamoto, and N. Asokan. Is this App Safe? A Large Scale Study on Application Permissions and Risk Signals. In *Proceedings of the 21st International Conference on World Wide Web*, 2012.

[10] Erika Chin, Adrienne Porter Felt, Vyas Sekar, and David Wagner. Measuring User Confidence in Smartphone Security and Privacy. SOUPS '12. Association for Computing Machinery, 2012. https://doi.org/10.1145/2335356.2335358.

[11] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Academic Press, 2013.

[12] Jayati Dev, Pablo Moriano, and L. Jean Camp. Lessons learnt from comparing WhatsApp privacy concerns across Saudi and Indian populations. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. USENIX Association, August 2020.

[13] Serge Egelman, Marian Harbach, and Eyal Peer. Behavior Ever Follows Intention? A Validation of the Security Behavior Intentions Scale (SeBIS). In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016.

[14] Serge Egelman and Eyal Peer. Scaling the Security Wall: Developing a Security Behavior Intentions Scale (SeBIS). In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2015. https://doi.org/10.1145/2702123.2702249.

[15] Pardis Emami Naeini, Henry Dixon, Yuvraj Agarwal, and Lorrie Cranor. Exploring How Privacy and Security Factor into IoT Device Purchase Behavior. In *CHI '19: Proceedings of CHI Conference on Human Factors in Computing Systems*, 2019. https://dl.acm.org/doi/10.1145/3290605.3300764.

[16] Cori Faklaris, Laura A. Dabbish, and Jason I. Hong. A Self-Report Measure of End-User Security Attitudes (SA-6). In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. USENIX Association, 2019. https://www.usenix.org/conference/soups2019/presentation/faklaris.

[17] Adrienne Porter Felt, Kate Greenwood, and David Wagner. The Effectiveness of Application Permissions. In *2nd USENIX Conference on Web Application Development (WebApps 11)*. USENIX Association, June 2011.

[18] Joseph A. Gliem and Rosemary R. Gliem. Calculating, Interpreting, and Reporting Cronbach's Alpha Reliability Coefficient for Likert-Type Scales. Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education, 2003.

[19] Thomas Groß. Validity and Reliability of the Scale Internet Users' Information Privacy Concern (IUIPC). PETs Symposium, 2020. https://petsymposium.org/2021/files/papers/issue2/popets-2021-0026.pdf.

[20] Friedrich M. Götz, Stefan Stieger, and Ulf-Dietrich Reips. Users of the main smartphone operating systems (iOS, Android) differ only little in personality. *PLOS ONE*, 2017. https://doi.org/10.1371/journal.pone.0176921.

[21] Marian Harbach, Alexander De Luca, Nathan Malkin, and Serge Egelman. *Keep on Lockin' in the Free World: A Multi-National Comparison of Smartphone Locking*. Association for Computing Machinery, 2016. https://doi.org/10.1145/2858036.2858273.

[22] Myles Hollander, Douglas A. Wolfe, and Eric Chicken. *Nonparametric Statistical Methods*, volume 751. John Wiley & Sons, 2013.

[23] Iulia Ion, Niharika Sachdeva, Ponnurangam Kumaraguru, and Srdjan Čapkun. Home is Safer than the Cloud! Privacy Concerns for Consumer Cloud Storage. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, SOUPS '11. Association for Computing Machinery, 2011. https://doi.org/10.1145/2078827.2078845.

[24] Ronald Kainda, Ivan Flechais, and Andrew William Roscoe. Usability and security of out-of-band channels in secure device pairing protocols. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, 2009.

[25] Ruogu Kang, Stephanie Brown, Laura Dabbish, and Sara Kiesler. Privacy Attitudes of Mechanical Turk Workers and the U.S. Public. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*. USENIX Association, 2014. https://www.usenix.org/conference/soups2014/proceedings/presentation/kang.

[26] Patrick Kelley. Privacy, measurably, isn't dead. USENIX Association, February 2021.

[27] Jennifer King. How Come I'm Allowing Strangers to Go Through My Phone? Smartphones and Privacy Expectations. *SSRN Electronic Journal*, 2012.

[28] Arun Kumar, Nitesh Saxena, Gene Tsudik, and Ersin Uzun. A comparative study of secure device pairing methods. *Pervasive and Mobile Computing*, 2009.

[29] Deepak Kumar, Riccardo Paccagnella, Paul Murley, Eric Hennenfent, Joshua Mason, Adam Bates, and Michael Bailey. Skill squatting attacks on Amazon Alexa. In *27th USENIX Security Symposium (USENIX Security 18)*. USENIX Association, August 2018.

[30] Ponnurangam Kumaraguru and Lorrie Cranor. Privacy indexes : A Survey of Westin's Studies. 2005. https://www.cs.cmu.edu/~ponguru/CMU-ISRI-05-138.pdf.

[31] Imane Lamiche, Guo Bin, Yao Jing, Zhiwen Yu, and Abdenour Hadid. A continuous smartphone authentication method based on gait patterns and keystroke dynamics. *Journal of Ambient Intelligence and Humanized Computing*, 2019.

[32] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, are you listening? Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), November 2018.

[33] Jialiu Lin, Bin Liu, Norman Sadeh, and Jason I. Hong. Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*. USENIX Association, July 2014.

[34] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi, Shikun (Aerin) Zhang, Norman Sadeh, Yuvraj Agarwal, and Alessandro Acquisti. Follow My Recommendations: A Personalized Privacy Assistant for Mobile App Permissions. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. USENIX Association, June 2016.

[35] Naresh K. Malhotra, Sung S. Kim, and James Agarwal. Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model. *Information Systems Research*, 2004. https://doi.org/10.1287/isre.1040.0032.

[36] Nathan Malkin, Joe Deatrick, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. Privacy Attitudes of Smart Speaker Users. *Proceedings on Privacy Enhancing Technologies*, 2019.

[37] Philipp Markert, Daniel V. Bailey, Maximilian Golla, Markus Dürmuth, and Adam J. Aviv. This pin can be easily guessed: Analyzing the security of smartphone unlock pins. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2020.

[38] Kristopher Micinski, Daniel Votipka, Rock Stevens, Nikolaos Kofinas, Michelle L. Mazurek, and Jeffrey S. Foster. User interactions and permission use on Android. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017.

[39] Alexios Mylonas, Anastasia Kastania, and Dimitris Gritzalis. Delegate the smartphone user? Security awareness in smartphone platforms. *Computers & Security*, 2013. https://doi.org/10.1016/j.cose.2012.11.004.

[40] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 2017.

[41] Adrian Perrig and Dawn Song. Hash visualization: A new technique to improve real-world security. In *International Workshop on Cryptographic Techniques and E-Commerce*, volume 25, 1999.

[42] Sören Preibusch. Guide to measuring privacy concern: Review of survey and observational instruments. *International Journal of Human-Computer Studies*, 2013.

[43] Lina Qiu, Alexander De Luca, Ildar Muslukhov, and Konstantin Beznosov. Towards understanding the link between age and smartphone authentication. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.

[44] Prashanth Rajivan and Jean Camp. Influence of privacy attitude and privacy cue framing on Android app choices. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. USENIX Association, 2016. https://www.usenix.org/conference/soups2016/workshop-program/wpi/presentation/rajivan.

[45] Elissa M. Redmiles. "Should I worry?" A cross-cultural examination of account security incident response. *CoRR*, abs/1808.08177, 2018. http://arxiv.org/abs/1808.08177.

[46] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How well do my results generalize? Comparing security and privacy survey results from mturk, web, and telephone samples. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019.

[47] Elissa M. Redmiles, Ziyun Zhu, Sean Kross, Dhruv Kuchhal, Tudor Dumitras, and Michelle L. Mazurek. Asking for a Friend: Evaluating Response Biases in Security User Studies. CCS '18. Association for Computing Machinery, 2018. https://doi.org/10.1145/3243734.3243740.

[48] Lena Reinfelder, Zinaida Benenson, and Freya Gassmann. Differences between Android and iPhone Users in Their Security and Privacy Awareness. In *Proceedings of the 11th International Conference on Trust, Privacy and Security in Digital Business*. Springer, 2014. https://link.springer.com/chapter/10.1007/978-3-319-09770-1_14.

[49] Raina Samuel, Philipp Markert, Adam J. Aviv, and Iulian Neamtiu. Knock, Knock. Who's There? On the security of LG's knock codes. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. USENIX Association, August 2020.

[50] Faysal Hossain Shezan, Hang Hu, Jiamin Wang, Gang Wang, and Yuan Tian. Read between the lines: An empirical measurement of sensitive applications of voice personal assistant systems. In *Proceedings of The Web Conference 2020*, 2020.

[51] Janice Sipior, Burke Ward, and Regina Connolly. Empirically assessing the continued applicability of the IUIPC construct. *Journal of Enterprise Information Management*, 26, 2013.

[52] H. Jeff Smith, Sandra J. Milberg, and Sandra J. Burke. Information Privacy: Measuring Individuals' Concerns about Organizational Practices. *MIS Quarterly*, 1996. http://www.jstor.org/stable/249477.

[53] Daniel Smullen, Yuanyuan Feng, Shikun Aerin Zhang, and Norman Sadeh. The best of both worlds: Mitigating trade-offs between accuracy and user burden in capturing mobile app privacy preferences. *Proceedings on Privacy Enhancing Technologies*, 2020.

[54] Peter Story, Daniel Smullen, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. From Intent to Action: Nudging Users Towards Secure Mobile Payments. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. USENIX Association, August 2020.

[55] Madiha Tabassum, Tomasz Kosiński, Alisa Frik, Nathan Malkin, Primal Wijesekera, Serge Egelman, and Heather Richter Lipford. Investigating users' preferences and expectations for always-listening voice assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, December.

[56] Madiha Tabassum, Tomasz Kosinski, and Heather Richter Lipford. "I don't own the data": End user perceptions of smart home device data practices and risks. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. USENIX Association, August 2019.

[57] Joshua Tan, Lujo Bauer, Joseph Bonneau, Lorrie Faith Cranor, Jeremy Thomas, and Blase Ur. Can unicorns help users compare crypto key fingerprints? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017.

[58] Prolific Team. Representative Samples FAQ. 2019. https://researcher-help.prolific.co/hc/en-gb/articles/360019238413-Representative-Samples-FAQ.

[59] Güliz Seray Tuncay, Jingyu Qian, and Carl A. Gunter. See No Evil: Phishing for Permissions with False Transparency. In *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association, August 2020.

[60] Sebastian Uellenbeck, Markus Dürmuth, Christopher Wolf, and Thorsten Holz. Quantifying the security of graphical passwords: The case of Android unlock patterns. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, 2013.

[61] Elham Vaziripour, Devon Howard, Jake Tyler, Mark O'Neill, Justin Wu, Kent Seamons, and Daniel Zappala. I Don't Even Have to Bother Them! Using Social Media to Automate the Authentication Ceremony in Secure Messaging. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.

[62] Elham Vaziripour, Justin Wu, Mark O'Neill, Daniel Metro, Josh Cockrell, Timothy Moffett, Jordan Whitehead, Nick Bonner, Kent Seamons, and Daniel Zappala. Action Needed! Helping Users Find and Complete the Authentication Ceremony in Signal. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*. USENIX Association, August 2018.

[63] Zack Whittaker. iOS 13: Here are the new security and privacy features you need to know. *TechCrunch*, 2019.

[64] Primal Wijesekera, Joel Reardon, Irwin Reyes, Lynn Tsai, Jung-Wei Chen, Nathan Good, David Wagner, Konstantin Beznosov, and Serge Egelman. Contextualizing privacy decisions for better prediction (and protection). In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[65] Allison Woodruff, Vasyl Pihur, Sunny Consolvo, Laura Brandimarte, and Alessandro Acquisti. Would a Privacy Fundamentalist Sell Their DNA for $1000 ... If Nothing Bad Happened as a Result? The Westin Categories, Behavioral Intentions, and Consequences. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*. USENIX Association, 2014. https://www.usenix.org/conference/soups2014/proceedings/presentation/woodruff.

[66] Justin Wu, Cyrus Gattrell, Devon Howard, Jake Tyler, Elham Vaziripour, Daniel Zappala, and Kent Seamons. "Something isn't secure, but I'm not sure how that translates into a problem": Promoting autonomy by designing for understanding in Signal. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, 2019.

[67] Eric Zeng, Shrirang Mare, and Franziska Roesner. End user security and privacy concerns with smart homes. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*. USENIX Association, July 2017.

[68] Miaoyi Zeng, Shuaifu Lin, and D. Armstrong. Are All Internet Users' Information Privacy Concerns (IUIPC) Created Equal? 2020. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1048&context=trr.

A Additional plots

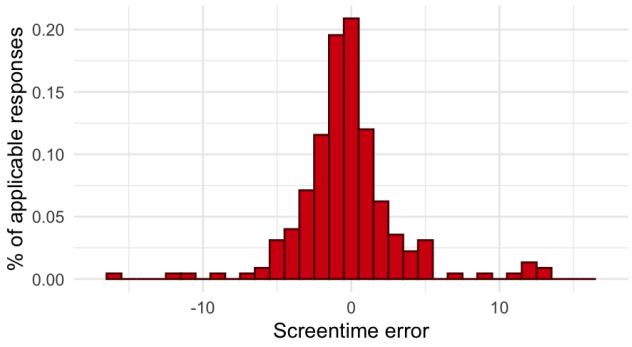
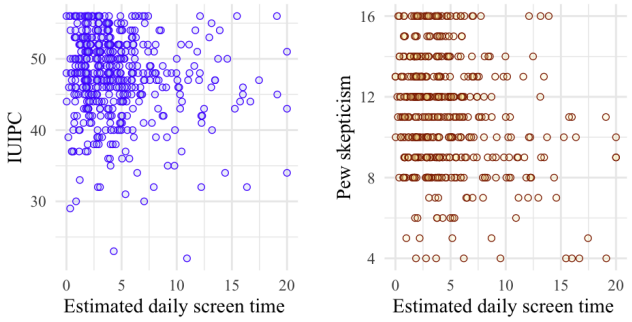
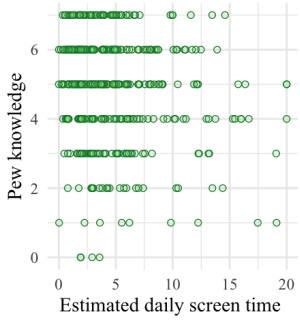


Figure 4: Histogram of differences in participant daily screen-time estimates vs. system screen-time report. The plot includes data from 225 (88 Android, 137 iOS) participants who provided both.



(a) IUIPC-8

(b) Pew skepticism



(c) Pew knowledge

Figure 5: Higher screen-time estimate was associated with lower IUIPC-8, Pew skepticism, and Pew knowledge scales.

B Survey Questionnaire

Consent and validation

1. *Consent form is shown, and consent is given*
2. In what country do you currently reside?
    - ○ United Kingdom
    - ○ United States
    - ○ Ireland
    - ○ Germany
    - ○ France
    - ○ Spain
    - ○ Other

    End survey if not United States
3. Please enter your Prolific ID here:
    [Free text]

Part 1: Device Background Screener Questions

1. How many smartphones do you currently own and use?
    - ○ 0
    - ○ 1
    - ○ 2+

    End survey if 0 is selected
2. For what purposes do you use your smartphone devices? Select all that apply. *[Displayed if "How many smartphones do you currently own and use?" 2+ Is Selected]*
    - ○ Personal
    - ○ Work
    - ○ Other: [Free text]
3. Please consider your PERSONAL smartphone device to be your primary device for the remainder of this survey. *[Displayed if "For what purposes do you use your smartphone devices? Select all that apply." Personal Is Selected AND "How many smartphones do you currently own and use?" 2+ Is Selected]*

    Page Break
4. Please consider your WORK smartphone device to be your primary device for the remainder of this survey. *[Displayed if "For what purposes do you use your smartphone devices? Select all that apply." Personal Is Not Selected AND "How many smartphones do you currently own and use?" 2+ Is Selected]*

    Page Break
5. How frequently do you use a voice assistant on your primary smartphone? (i.e. Hey Siri, OK Google, etc.)
    - ○ Multiple times a day
    - ○ Almost once a day
    - ○ A few times a week
    - ○ A few times a month
    - ○ Almost never
6. Which Operating System do you use for your primary smartphone?
    - ○ iOS
    - ○ Android
    - ○ Windows
    - ○ Other or Not Applicable [Free text]
    - ○ I don't know

    End survey if not iOS or Android
7. Which of the following smart device(s) do you currently own?
    - ○ Smart Speaker
    - ○ Smart Doorbell
    - ○ Smart Thermostat
    - ○ Smart TV
    - ○ Smart Fridge
    - ○ None of the above
    - ○ Other: [Free text]
8. Which smart speaker (voice assistant) do you use? Select all that apply. *[Displayed if "Which of the following smart device(s) do you currently own?" Smart Speaker Is Selected]*
    - ○ Echo Device
    - ○ Google Home
    - ○ Apple HomePod
    - ○ Other: [Free text]
9. How frequently do you use your smart speaker(s)? (i.e. an Echo Device, Google Home, etc.) *[Displayed if "Which of the following smart device(s) do you currently own?" Smart Speaker Is Selected]*
    - ○ Multiple times a day
    - ○ Almost once a day
    - ○ A few times a week
    - ○ A few times a month
    - ○ Almost never

10. How much time do you spend on your primary smartphone on average?
    - Please estimate your daily average in hours. [Slider]

Page Break

11. Do you currently have access to your primary smartphone device? We may ask you to refer to your smartphone during this survey.
    - Yes
    - No

Part 2A: iOS Background Questions

[Displayed if "Which Operating System do you use for your primary smartphone?" iOS Is Selected]

1. What is the iPhone model of your primary smartphone?
    - SE (1st or 2nd generation)
    - 12, 12 Pro, 12 Mini
    - 11, 11 Pro, or 11 Pro Max
    - X, XS, XS Max, or XR
    - 8 or 8 Plus
    - 7 or 7 Plus
    - 6, 6 Plus, 6S, or 6S Plus
    - 5, 5S, or 5C
    - Other: [Free text]
    - I don't know
2. Is your primary smartphone device jailbroken?
    - Yes
    - No
    - I don't know
3. Please navigate through the following steps on your smartphone to answer the following question accurately:
   Settings App > General > About > Software Version
   Which version of iOS does your primary smartphone have? *[Displayed if "Do you currently have access to your primary smartphone device? We may ask you to refer to your smartphone during this survey." Yes Is Selected]*
    - iOS 14
    - iOS 13
    - iOS 12
    - iOS 11
    - iOS 10
    - iOS 9
    - Other: [Free text]

Page Break

4. Please navigate through the following steps on your smartphone to answer the following question accurately:
   Settings > Screen Time
   Can you see your daily average in screen time for the past week? Note that some phones show daily numbers but don't show an average, please select no if that's the case. *[Displayed if "Do you currently have access to your primary smartphone device? We may ask you to refer to your smartphone during this survey." Yes Is Selected]*
    - Yes
    - No

Page Break

5. How much time on average do you spend on your primary smartphone? Please report your daily average. *[Displayed if "Please navigate through the following steps on your smartphone to answer the following question accurately:*
   Settings > Screen Time
   Can you see your daily average in screen time for the past week? Note that some phones show daily numbers but don't show an average, please select no if that's the case" Yes Is Selected]
    - hour(s) [Free text]
    - minute(s) [Free text]

Part 2B: Android Background Questions

[Displayed if "Which Operating System do you use for your primary smartphone?" Android Is Selected]

1. What is the Android model of your primary smartphone?
    - Blackberry
    - HTC
    - Lenovo
    - LG
    - Motorola
    - Nexus
    - Nokia
    - Google Pixel

- Samsung Galaxy
- Sony Xperia
- Other: [Free text]

2. Is your primary smartphone device rooted?
   - My device is rooted
   - My device is non-rooted
   - I don't know

3. Please navigate through the following steps on your smartphone to answer the following question accurately:
   Settings App > About Phone > Android Version Settings App > About Phone > Software Information > Android Version
   Which version of Android does your primary smartphone have?
   - Android 11
   - Android 10
   - Pie 9.0
   - Oreo 8.0-8.1
   - Nougat 7.0-7.1.2
   - Marshmallow 6.0-6.0.1
   - Lollipop 5.0-5.1.1
   - KitKat 4.4-4.4.4
   - Other: [Free text]

Page Break

4. Please navigate through the following steps on your smartphone to answer the following question accurately:
   Settings > Digital Wellbeing
   Can you see your daily average in screen time for the past week? Note that some phones show daily numbers but don't show an average, please select no if that's the case. *[Displayed if "Do you currently have access to your primary smartphone device? We may ask you to refer to your smartphone during this survey." Yes Is Selected]*
   - Yes
   - No

Page Break

5. How much time do you spend on your primary smartphone on average? Please report your daily average. *[Displayed if "Please navigate through the following steps on your smartphone to answer the following question accurately:*
   Settings > Digital Wellbeing
   Can you see your daily average in screen time for the past week? Note that some phones show daily numbers but don't show an average, please select no if that's the case." Yes Is Selected]
   - hour(s) [Free text]
   - minute(s) [Free text]

Part 3: Pew Knowledge Questions

1. What does it mean when a website has "https://" at the beginning of its URL, as opposed to "http://" without the "s"?
   - Information entered into the site is encrypted
   - The content on the site is safe for children
   - The site is only accessible to people in certain countries
   - The site has been verified as trustworthy
   - Not sure

2. Many web browsers offer a feature known as "private browsing" or "incognito mode." If someone opens a webpage on their computer at work using incognito mode, which of the following groups will NOT be able to see their online activities?
   - The group that runs their company's internal computer network
   - Their company's internet service provider
   - A coworker who uses the same computer
   - The websites they visit while in private browsing mode
   - Not sure

3. When a website has a privacy policy, it means that the site…
   - Has created a contract between itself and its users about how it will use their data
   - Will not share its users' personal information with third parties
   - Adheres to federal guidelines about deceptive advertising practices
   - Does not retain any personally identifying information about its users
   - Not sure

4. If a website uses cookies, it means that the site ...
    - Can see the content of all the files on the device you are using
    - Is not a risk to infect your device with a computer virus
    - Will automatically prompt you to update your web browser software if it is out of date
    - Can track your visits and activity on the site
    - Not sure
5. Which of the following is the largest source of revenue for most major social media platforms?
    - Exclusive licensing deals with internet service providers and cellphone manufacturers
    - Allowing companies to purchase advertisements on their platforms
    - Hosting conferences for social media influencers
    - Providing consulting services to corporate clients
    - Not sure
6. Where might someone encounter a phishing scam?
    - In an email
    - On social media
    - In a text message
    - On a website
    - All of the above
    - None of the above
    - Not sure
7. The term "net neutrality" describes the principle that ...
    - Internet service providers should treat all traffic on their networks equally
    - Social media platforms must give equal visibility to conservative and liberal points of view
    - Online advertisers cannot post ads for housing or jobs that are only visible to people of a certain race
    - The government cannot censor online speech
    - Not sure

Part 4: SA-6

1. Please rate your agreement or disagreement with the following statements. *Options: {Strongly agree, Agree, Neutral, Disagree, Strongly disagree}*
    - I seek out opportunities to learn about security measures that are relevant to me.
    - I am extremely motivated to take all the steps needed to keep my online data and accounts safe.
    - Generally, I diligently follow a routine about security practices.
    - I often am interested in articles about security threats.
    - I always pay attention to experts' advice about the steps I need to take to keep my online data and accounts safe.
    - I am extremely knowledgeable about all the steps needed to keep my online data and accounts safe.

Page Break

2. You answered you [participant's selected answer] with the following statement: I often am interested in articles about security threats. Why did you feel this way? Please explain why you chose this answer. [Free text] *(Used as an attention check)*

Part 5: IUIPC

Please rate your agreement or disagreement with the following statements. *Options: {Strongly agree, Agree, Somewhat agree, Neutral, Somewhat disagree, Disagree, Strongly disagree}*

1. **Control**
    - Consumer online privacy is really a matter of consumers' right to exercise control and autonomy over decisions about how their information is collected, used, and shared.
    - Consumer control of personal information lies at the heart of consumer privacy.
    - I believe that online privacy is invaded when control is lost or unwillingly reduced as a result of a marketing transaction.
2. **Awareness**
    - Companies seeking information online should disclose the way the data are collected, processed, and used.
    - A good consumer online privacy policy should have a clear and conspicuous disclosure.
    - It is very important to me that I am aware and knowledgeable about how my personal information will be used.
3. **Collection**
    - It usually bothers me when online companies ask me for personal information.

- When online companies ask me for personal information, I sometimes think twice before providing it.
- It bothers me to give personal information to so many online companies.
- I'm concerned that online companies are collecting too much personal information about me.

Part 6: Pew Company Questions

1. How confident are you, if at all, that companies will do the following things? *Options: {Very confident, Somewhat confident, Not too confident, Not confident at all}*
   - Follow what their privacy policies say they will do with your personal information
   - Promptly notify you if your personal data has been misused or compromised
   - Publicly admit mistakes and take responsibility when they misuse or compromise their users' personal data
   - Use your personal information in ways you will feel comfortable with
   - Be held accountable by the government if they misuse or compromise your data

Page Break

2. You answered you are [participant's selected response] that companies will: Publicly admit mistakes and take responsibility when they misuse or compromise their users' personal data. Why did you feel this way? Please explain why you chose this answer. [Free text] *(Used as an attention check)*

Part 7: How Well Do My Results Generalize? (as it appears in [46])

1. Do you feel as you already know enough about . . . *Options: {Already know enough, Would like to learn more, Does not apply, Do not know}*
   - Choosing strong passwords to protect your online accounts
   - Managing the privacy settings for the information you share online
   - Understanding the privacy policies of the websites and applications you use
   - Protecting the security of your devices when using public Wifi networks
   - Protecting your computer or mobile devices from viruses and malware
   - Avoiding online scams and fraudulent requests for your personal information

Part 8: Demographics

1. Please indicate your age. If you'd prefer not to answer, you can skip this question.
   - Use the slider to indicate your age. [Slider]
2. What gender do you best identify with?
   - Man
   - Woman
   - Non-binary
   - Prefer to self-describe [Free text]
   - Prefer not to answer
3. Which of the following best describes your race? Select all that apply.
   - White
   - Black or African American
   - American Indian or Alaska Native
   - Hispanic or Latino
   - Asian
   - Native Hawaiian or Pacific Islander
   - Other [Free text]
   - Prefer not to answer
4. Please specify the highest degree of level of school you have completed or currently attending.
   - No high school degree
   - High school graduate, diploma or the equivalent (for example, GED)
   - Some college credit, no degree
   - Trade, technical, vocational training
   - Associate's degree
   - Bachelor's degree
   - Master's degree
   - Professional degree
   - Doctorate degree
   - Other [Free text]
   - Prefer not to answer
5. What is your current employment status?
   - Employed Full-Time
   - Employed Part-Time
   - Self-employed

- ○ Unemployed
- ○ Student
- ○ Home-maker
- ○ Retired
- ○ Disabled
- ○ Prefer not to answer

6. What is your annual household income?

- ○ Up to $25,000
- ○ $25,000 to $49,999
- ○ $50,000 to $74,999
- ○ $75,000 to $99,999
- ○ $100,000 or more
- ○ Prefer not to answer

7. How frequently do you give computer or technology advice (e.g., to friends, family, or colleagues)?

- ○ Almost always
- ○ Often
- ○ Sometimes
- ○ Rarely
- ○ Never

end of survey

"How I Know For Sure": People's Perspectives on Solely Automated Decision-Making (SADM)

Smirity Kaushik[1], Yaxing Yao[2], Pierre Dewitte[3], Yang Wang[1]
[1]University of Illinois at Urbana-Champaign [2]University of Maryland, Baltimore County
[3]Katholieke Universiteit Leuven Centre for IT & IP
{smirity2, yvw}@illinois.edu, {yaxingya}@umbc.edu, {pierre.dewitte}@kuleuven.be

Abstract

Algorithms are used to make automated decisions that can affect individuals in numerous domains. The General Data Protection Regulation (GDPR) of the European Union (EU) has granted citizens some rights regarding solely automated decision-making (SADM), including obtaining an explanation of such processing. It is unclear, however, how organizations should support people in effectively exercising such rights. We conducted an online survey to understand people's perspectives on SADM. We found that our respondents had several misunderstandings about the SADM right, such as opt-out of SADM ahead of time. We also identified various attributes of SADM that our respondents desired to understand, including new attributes (e.g., actionable information about what they can practically do to improve future decision outcomes) not covered by implementation guidelines of the GDPR. Our respondents also anticipated many challenges with SADM, including not knowing when SADM is applied to them. We discuss design implications of our results on how to support people in coping with SADM, for instance, the design of icons to represent SADM processing and explanation templates that cover a common set of attributes and can be personalized to explain a specific SADM decision.

1 Introduction

From job applications to insurance premiums to targeted ads, algorithms have increasingly been used to make automated decisions that can affect individuals [16,66]. While using algorithms to automatically make decisions about individuals may allow companies to increase efficiency and save resources, they also pose significant threats to individuals and society such as privacy violations, social segregation, discrimination, and unjustified denials of service [4, 25, 28, 60, 62]. Furthermore, these automated systems often remain opaque to public scrutiny and understanding, thus leading to a lack of decision acceptance and trustworthiness in these algorithmic practices.

The SADM right. One key aspect of making automated decision-making fair, accountable, and transparent is to provide sensible explanations of these algorithmic decisions to individuals who might be affected. For instance, the European Union (EU) General Data Protection Regulation (GDPR), which entered into force on May 25 2018, defines such an algorithmic decision as "decision based solely on automated processing, including profiling, which produces legal effects concerning the data subject or similarly significantly affects him or her" [70]. Among other citizen rights, the law allows citizens (1) to obtain an explanation of the logic involved as well as the significance and the envisaged consequence of such processing, and (2) to request human intervention, express a point of view, and contest the decision [70]. Since these rules are closely related and to simplify the reference to these rules, we coined an umbrella term "Right against solely automated decision-making" (or the *SADM right*).

GDPR background. The GDPR applies to (1) companies that have an establishment in the EU and (2) companies not based in the EU but offer goods and services to people living in the EU or monitor their behavior if that behavior occurs in the EU [70]. For instance, if an American organization (e.g., company or university) has employees, regardless of their citizenship, who live in the EU (e.g., university staff who work in the study abroad program in France), then the US organization needs to comply with GDPR for those employees. The GDPR and its underlying principles have also substantially impacted other privacy legislation around the world. For instance, Brazil and India are adopting GDPR-like legislation. The UK enacted EU GDPR's requirements into a UK law (Data Protection Act 2018 [17]) and later amended it in 2019 to form a new UK-specific data protec-

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

tion regulation (UK-GDPR [18]) for post-Brexit transition. The UK has also provided guidelines for explaining AI-based decisions or AI-assisted decisions where humans and AI are both involved [69]. In part inspired by the GDPR, the US has also taken significant steps to improve transparency and consumer privacy with state legislations such as California Consumer Privacy Act (2018) [56], California Privacy Rights Act (2020) [52] and proposals of an omnibus federal privacy law [45].

Research motivations. Despite the potential benefits of GDPR and the SADM right more specifically, it is unclear how people perceive SADM or how organizations should support people in exercising this right. This is in part because GDPR was not very specific about this right, which is still open to interpretation regarding its implementation. It is also why the EU A29WP Working Party (now replaced by the European Data Protection Board) has been creating more actionable guidelines on SADM [26]. Moreover, it is increasingly recognized that understanding citizens' perspectives on SADM is needed. For instance, both legal and HCI scholars [6, 29, 38, 55, 57, 63] have advocated for seeking citizens' inputs as crucial and timely in informing the refinement of these actionable guidelines and the ways in which companies can effectively support citizens in exercising the SADM right. Despite the need and importance of understanding citizens' perspectives on the SADM right, there is a lack of empirical research on this topic. Our research aims to help fill this gap.

Research questions. Whether organizations can provide people an effective explanation of their SADM practices is a crucial element of their ethical use of algorithms and cultivating consumer trust in SADM systems. However, what people consider as constituents of an effective explanation of SADM is an open question. Given the broad impact of GDPR on algorithms, people, and ethics, our study contributes to the understanding of people's expectations of the SADM right and the design of socio-technical mechanisms that empower people to exercise this right. Our long-term goal is to design such mechanisms. To inform such design, this paper focuses on three research questions:

- **RQ1:** What are people's understandings of SADM right?
- **RQ2:** What aspects of SADM do people want to know?
- **RQ3:** What challenges do people anticipate in exercising the SADM right?

RQ1 can elucidate how people perceive and expect to exercise the SADM right. RQ2 can provide a baseline list of aspects that companies can consider including in the explanations of their SADM practices. RQ3 can identify peoples' anticipated challenges of exercising the SADM right. Together, answering these questions can inform future design to support people's needs regarding SADM.

Study and findings. To answer our research questions, we conducted an online survey with 392 respondents from the UK and the US. Our research approach gives voice to ordinary citizens' perspectives. It is akin to governments conducting "citizen juries" that directly inquire and incorporate people's inputs on important public policy issues (e.g., [68]). We deliberately chose one country (UK) where the SADM right is directly supported while another (US) that currently does not. It could help us to explore whether they would have drastically different views on SADM. This was found not the case in our study, implying that our respondents shared common expectations of SADM systems and their explanations.

For RQ1, our respondents had many misunderstandings about the SADM right, e.g., incorrectly assuming that the right allows them to opt out of SADM ahead of time or that they could deny the use of personal information for SADM processing. For RQ2, our respondents desired to know more about SADM than what policy-makers have suggested organizations provide (e.g., type of information, source of information, logic used). For instance, our respondents expected to receive personalized explanations including factors considered. It indicated a stronger need of the respondents to seek justification for the decisions made, especially the negative ones. For RQ3, our respondents anticipated a wide variety of challenges. Some of these were unique to SADM, such as the difficulty for people to know when they are subject to SADM.

Research contributions. This research makes two primary contributions. First, it has many novel empirical results on people's perspectives on the SADM and the related right. Specifically, it uncovers people's misunderstandings of the right, which will hinder their effective exercise of the right to protect themselves. The study also identifies attributes of SADM that people want to understand beyond what has been recommended by policymakers. Our research also uncovers some unique challenges anticipated by our participants to exercise the SADM right. Second, we propose several design implications, based on our study results, for organizations to support citizens exercise this right. For instance, designing and using (standardized) icons to represent SADM processing, personalized explanation templates to explain SADM outcomes, and SADM sandboxes that allow users to explore the SADM systems to improve future decisions about them.

2 Related Work

In this section, we present the relevant literature on people's perceptions of algorithmic accountability, fairness, methods of creating explanations of AI systems, and people's challenges with legal concepts (with a focus on privacy policies).

2.1 Perceptions of Algorithmic Accountability

Algorithmic accountability has recently received a lot of traction with the increased use of algorithms in high-impact domains and the changes in the regulatory landscape, such as the

implementation of the EU's GDPR. Bovens describe accountability as 'a relationship between an actor and a forum, in which an actor must explain and to justify his or her conduct, the forum can pose questions and pass judgment, and the actor may face consequences' [7]. Recent scholarship has conceptualized algorithmic accountability, building on the traditional accountability literature. For instance, Wieringa [72] adapted Bovens's [7] widely accepted definition of accountability in the context of algorithmic accountability. The actors (i.e., algorithm developers, decision-makers) are accountable to explain or justify the algorithmic decisions (account) in a forum that can question these actors, often with consequences [72]. Additionally, it requires a perspective, i.e., identifying what needs to be accounted for in the algorithmic system [72]. For instance, Coglianese and Lehr [14] distinguished actors by their roles to determine the appropriate actor for a particular situation. They noted that it is important to identify 'who within an agency actually wields algorithm-specifying power' [14]. According to Kolkman [30], a forum provides a platform for users to understand the decision-making process and to engage with it. It can further impose consequences on the actors. The GDPR [1] through its SADM right provides that forum to an individual citizen in the context of algorithmic accountability and further imposes legal accountability (consequences) for actors. From a technical perspective, Kroll et al. [41] linked algorithmic accountability to a system's life cycle, identifying two possible approaches to algorithmic accountability: ex-ante (before the decision is made) and ex-post (after the decision is made). Neyland [51] argued that accountability should be considered throughout different stages of algorithmic system, i.e., design, implementation, and evaluation, as accountability is a shared responsibility between designers, developers, and users throughout the system's life cycle. Finally, several researchers [10, 13, 14] also argued that algorithmic accountability has a direct relationship to the measure of human involvement. For instance, in case of a human-out-of-the-loop system (or a SADM system), the degree of accountability increases manifolds as there is no human oversight. Our work was partly motivated by the notion of algorithmic accountability, which underpins the SADM right, making designers of SADM systems responsible for explaining and justifying their systems and practices.

2.2 Perceptions of Algorithmic Fairness

The concept of fairness is vast and ambiguous and is used differently across disciplines [48]. Fairness broadly refers to an equitable outcome that can be justified reasonably for a purpose within a context or domain. It further includes the dimension of unfairness which elaborates on what and who is considered capable of violating fairness. Fairness also seeks to clarify who is to be protected and where such protection can be operationalized [48]. Recent work has explored algorithmic fairness in different ways such as building fair decision-making algorithms, determining peoples' perceptions of fairness in these systems. For this paper, we explored prior literature focusing on people's perceptions of algorithmic fairness in general and in specific application domains. For instance, Grgic-Hlaca et al. examined why people perceive the use of certain features as unfair in making decisions about individuals in general [31]. They proposed that people's unfairness concerns are multi-dimensional, based on various aspects such as the relevance, volitionality, and reliability of decision and moral judgment. Woodruff et al. explored the impact of algorithmic bias on marginalized groups based on demographic features such as race [73]. Other scholars have analyzed people's perceptions of algorithmic fairness in specific domains, such as real estate and finance. For instance, Lee and Baykal investigated people's perceptions of fair division algorithms (e.g., those designed to divide rent among tenants) compared to discussion-based group decision-making methods [42]. They found that participants perceived the algorithmic decisions to be less fair than group-based decisions because the former did not account for people's social behavior. Saxena et al. investigated ordinary people's attitude toward three notions of individual fairness in the context of loan decisions [61]. They found that people tend to prefer calibrated fairness which selects individuals in proportion to their merit.

Prior research has also examined algorithmic fairness from experts' perspectives. For instance, Veale et al. interviewed public sector machine learning practitioners regarding the challenges of incorporating public values into their work [71]. They found a disconnect between organizational realities and current research into algorithmic fairness. They proposed incorporating domain knowledge by designing usable privacy tools aimed at private sector managers and public sector bureaucrats. Similarly, Holstein et al. conducted a systematic investigation of commercial product teams' challenges in developing fairer machine learning systems [34]. It highlights the disconnection between the challenges faced by teams in practice and the proposed solutions in the literature review. Research suggests that people care about the fairness of algorithms as well as potential discrimination and biases when companies make decisions about them. Our study helps to understand whether people would consider fairness issues such as biases in the context of SADM.

2.3 Transparent and Explainable AI

The algorithmic black box makes it difficult for users to know how an algorithm works, mainly because the information is either of a certain level of secrecy or intellectual property or too complicated for users to understand [64]. The principle of transparency is related to such a black box. It refers to provide people with the details of knowledge/information that a system gains from its users implicitly. Such details may include how the service works, the potential consequences and

other types of data management (e.g., sensible data) [2]. For example, in the 2016 US Presidential Election, algorithmic transparency on Facebook became a key issue to "end the profiling" [11]. Research has been arguing that practicing the principle of transparency may have a significant impact on people's knowledge and behaviors. Lee and Boynton stated that people are more like to use a system properly and form a sense of trust toward the system designer and developers if they understand how the systems work [43]. The notion of transparency is very relevant to our present work because explanation of SADM is a form of transparency.

Improving transparency of AI systems has been an active area of research in the AI and machine learning community. There is a growing body of work on creating explanations of AI systems or decisions/predictions to improve their transparency. Hu et al. presents a survey of existing methods [35], which mainly differ by two dimensions: scope of the explanation (global vs. local) and how the explanation is generated (intrinsic vs. post-hoc). Global explanations focus on how the whole model/system works, whereas local explanations describe specific decisions made by the model. For instance, 'model-agnostic' approaches such as LIME (Local Interpretable Model-Agnostic Explanations) explain a specific prediction of an algorithm by learning a local model around that predicted value [59]. However, these approaches are primarily helpful in supporting experts (e.g., data analysts) [9]. Intrinsic explanations are built-in part of the model. For instance, decision tree rules [44] are a built-in aspect of the decision tree, which lends itself to easy human interpretation. In comparison, deep learning models (artificial neural networks) are often difficult to explain the logic behind these models. Thus, researchers have created (post-hoc) explanations after these models have been built.

From a policy perspective, Doshi-Velez and Kortz argue that an explanation of an algorithm for end-users requires similar level of accountability that is ascribed to the human decision-makers. This may be achieved by applying certain technical considerations such as using local explanations to reach the outcome without divulging company trade-secrets [21]. Selbst and Barcos [63] explore the use of existing laws to fix the interpretability challenge of machine learning algorithms by focusing on not only their logic but also their fairness. From an HCI perspective, Binns et al. argue that end-users expect system explanations to be similar to those from human decision makers [5]. This is because the end-users apply similar perceptions of justice to the context of automated decision-making that is applicable in human decision-making [5]. Eslami et al. conducted a qualitative study to explore end-users' perceptions of personal information used to make targeted ads [24]. They found that increased visibility of inferences made about users can lead to "algorithm disillusionment," the idea that users realize the limitations of those algorithms which they thought to be perfect before [24]. In another study, Kizilcec [40] found that over-explanation i.e., supplementing the procedural explanation of the grade-adjusting algorithm with the outcome-specific information to increase the transparency of the algorithm further led to students' distrust in the system. There should be a balance between lack of explanation and over-explanation of the algorithm process to cultivate user trust.

2.4 People's Challenges with Privacy Policies

Since the SADM right has legal meanings, we also looked at prior research that shows how ordinary people may struggle with legal concepts and documents. In the domain of privacy, these issues have been around for a long time [12]. For instance,people struggle with privacy policies, a form of explanation that describes an organization's data/privacy practices and are required in the US. Grossklags and Goods point out that privacy policies are often unstructured, jargon-filled, and thus difficult to read and comprehend [32]. Prior work has also shown less than 1% of the general population read these documents [3].

2.5 Summary

These lines of work suggest that people may have multi-dimensional perspectives on automated decision-making, and they may encounter challenges with legal concepts. The prior literature also calls for research on human-centered perspectives on SADM to provide accountable, fair, and transparent, yet balanced explanation to automated decision-making. Our study helps fill the gap by investigating people's perspectives of SADM, such as their understandings and expectations of the SADM right, the kinds of attributes of automated decision-making that they wish to understand, and what challenges they anticipate in exercising the SADM right.

3 Method

To answer our research questions, we conducted a large-scale online survey with respondents from the UK and US to understand common understandings and perceptions of automated decision-making. We chose surveys over interviews because the former allowed us to study a much larger sample and identify common patterns in their perceptions. Our IRB approved this research.

We decided to focus on the UK and US in this study for many reasons. First, while the UK needs to be compliant with UK-GDPR [18], the US currently does not. However, the underlying transparency, accountability, and fairness principles (i.e., providing transparency about companies' data practices) are much more broadly supported. For example, the Fair Information Practices principles that undergird privacy-related legislation in the US include an openness principle (organizations should be open about "developments, practices and

policies with respect to personal data") and an individual participation principle (individuals can access data about them or confirm whether an organization has data about them) [20]. SADM is a concrete type of data practice. We deliberately chose one country where the SADM right is directly supported while another that has not to see whether they would have drastically different perspectives on the idea of SADM. Second, the US is the world's leading Internet economy and the home for most Internet giants that provide services to global users. The UK is the largest internet economy in the G20 [17] surpassing other European countries such as Germany and France. Third, the two countries are from two continents and might have different cultures. Lastly, English is the official language of both countries allowing us to use the same survey.

3.1 A Video Tutorial of The SADM Right

Since our respondents might not know the SADM right, we created a short video with animations to educate them about the right[1]. We chose this format over others because a video with animations, an audio track, and text captions can be more engaging and accessible than text or audio alone. Below is a summary of how we designed this video tutorial.

First, we searched descriptions and examples of the SADM right from reliable resources such as the GDPR website [70], UK-ICO (UK Information Commissioner's Office) [53], CNIL (French Data Protection Authority) [50], and EU Working Party-29 (AWP29) Guidelines [26]. We specifically sought: (1) introduction of the right as part of the GDPR; (2) explanation of salient features of the right (e.g., under the SADM right, the users were introduced to the terms profiling and automated decision-making, along with an example of each term); and (3) a real-life example illustrating how the right can be exercised. We included a bank-related example from the EU's official website [15]. The example reads, *"You use an online bank for a loan. You are asked to insert your data and the bank's algorithm tells you whether the bank will grant you the loan or not and gives the suggested interest rate. You must be informed that you may express your opinion, contest the decision and demand that the decision made via the algorithm be reviewed by a person. Additionally, the company must also explain to you how the automated decision-making algorithm makes a decision about you and its envisaged legal or significant consequences for you."*

Second, the descriptions and examples were then integrated into a script (written in English) by one researcher, based on which another researcher created various graphical animations using Adobe Illustrator, Adobe Photoshop and iMovie. A third researcher, a legal scholar with expertise in the GDPR, reviewed the script and the animations to ensure their correctness and neutrality. When inaccuracies were identified, we updated the script and the animations accordingly. We performed this process iteratively until all researchers agreed on both the script and animations.

Third, we audio-recorded the final script in English, embedded the audio in the animations, and generated the video. The video was then incorporated into the survey. Before conducting the actual study in Fall 2018, we did a pilot on Amazon Mechanical Turk (US participants) and Prolific (UK participants) with the initial survey design. We received some feedback that the voice of the video was too monotonic. To address this issue, we on-boarded a broadcasting professional to re-record the audio track with more tones and variations to make the audio more engaging.

3.2 Survey Flow

Our survey had a total of 21 questions, including both open-ended and multiple-choice questions. We started by asking people's understanding of the SADM right. The first question asked whether respondents had heard of this right before, and the second asked about their understanding of the right. We then asked another two questions about their understanding of 'user profiling' and 'automated decision-making,' respectively. Next, respondents were asked to watch our video tutorial on the SADM right. Then, we asked a simple attention checking question ("In the video, did you spot a computer screen?") to ensure they paid attention to the video. We then asked them two multiple-choice questions (MCQ) to examine their understandings of the right, e.g., "which of the following statements best explains the Right against solely automated decision making?", "which of the following is an example of the Right against solely automated decision making?" For each MCQ question in the survey, they can select all/multiple responses that apply. However, out of the five answer options there was only one correct answer option to each of the two multiple-choice question, based on the legal definition of the right. For instance, the statement that best explained the SADM right was, *"It allows users to request companies to explain how the AI makes automated decisions about them."* Similarly, the correct example of SADM right was, *"An individual stumbled upon the social media settings page that shows an automatically generated profile that the company uses to provide a personalized news feed and other targeted ads. This person requests the company to explain the process of automatic profiling."* We then asked an open-ended question about decisions that may affect them significantly.

Next, to understand respondents' expectations about attributes to explain for the SADM right, we asked an open-ended question, "Regarding the example from the video about automatically making decisions on bank loans, what aspects of automated decision making would you like to be explained?" This was followed by a multiple-choice question (MCQ) asking respondents to choose attributes they would like to be included in the explanations about SADM. They can choose multiple responses from a randomized list of six

[1] https://youtu.be/BrQMqmPEWQs

pre-defined attributes, adapted from concrete examples in the 'Guidelines on Automated Decision-Making' created by the EU Working Party (A29WP [26]). These attributes include:

- *Type of information:* "Company must inform me what types of data (e.g., my name, age, address) are used in making automatic decisions about me."
- *Explaining logic involved:* "Company must inform me how data is used or what algorithm or method is used in making automatic decisions about me."
- *Fairness of algorithm:* "Company must inform me how it ensures that the algorithm or method used in making automatic decisions about me is fair and unbiased."
- *Source of information:* "Company must inform me where or how the company found/obtained my data that is used in making automatic decisions about me."
- *Company infrastructure:* "Company must provide me customer care services to contest automated decision making."
- *Data accuracy and updates:* "Company must update both my data and the algorithm for accurate decision making as well as inform me about these updates."

We then asked respondents to recall whether they have encountered an incident where they could have exercised this right. If they answered "Yes," we then asked them to provide details about that incident, how they could have exercised the right, what benefits and challenges they could have if they try to exercise the right, and any possible solution to address the challenges. If they answered "No," we directly asked them about the benefits, challenges, and solutions. Lastly, we asked their overall understanding of the right using a Likert scale question followed by an open-ended question.

3.3 Data Analysis

For the open-ended questions, we conducted a thematic analysis [8], a standard method for analyzing qualitative data. We used a software called Dedoose to code the open-ended survey responses qualitatively. Three researchers coded together and discussed a 10% subset of the data. Once the coders achieved a good understanding of the data, the three coders continued to code another 10% subset of the data independently, then discussed and reconciled their coding to develop the initial codebook. The inter-coder reliability is 0.89 (Cohen's Kappa), which is considered good [27]. Due to a large amount of qualitative data from our survey, we decided to involve another two research assistants to help with the coding process. The inter-coder reliability for the additional two coders was 0.88 (Cohen's Kappa). Using the agreed-upon code-book, one researcher and the two new coders repeated the procedure mentioned above to ensure that all coders shared a good understanding of the coding process. Then each of the five coders independently coded a subset of the rest of the data. We added new codes to the code-book when existing codes cannot capture the data. Upon completion, the final code-book contained over 300 lower-level codes. We then grouped all codes into higher-level themes, such as people's understanding or misunderstandings of the SADM right before and after watching the video (i.e., *understanding* questions), attributes of SADM that people desire to know (i.e., *attribute* questions), and different types of expected challenges in exercising the right (i.e., *challenge* questions).

Furthermore, we compared how UK and US respondents answered these key questions on understanding, attribute, and challenges regarding the percentage of respondents who selected each answer in the multiple-choice questions or mentioned a theme in the open-ended questions. Since each respondent can select/mention multiple answers in each question, we cannot perform Chi-Square tests comparing the two groups because the data is not independent (e.g., two answers were chosen by the same respondent). Instead, we treated each answer/theme as a separate, binary question. A respondent can only select/mention an answer or not. We then conducted tests of proportions between the UK and US respondents for each answer. Since these are essentially post-hoc comparisons, we applied Bonferroni correction to adjust the family-wise p value. Specifically, we divided the common 0.05 p value threshold by the number of answers/themes in each question. The adjusted p value threshold for statistical significance ranges from 0.007 to 0.01 depending on the specific questions

3.4 Participants

Similar to prior research, we used Amazon Mechanical Turk (MTurk, based in the US) to recruit US respondents (e.g., [39]) and Prolific (based in the UK) to recruit UK respondents (e.g., [54]). We could not recruit enough UK respondents from MTurk because MTurk is much less popular than Prolific in the UK. Similarly, we recruited US respondents on MTurk because it is much more popular than Prolific in the US. Therefore, we had to use two platforms, each for respondents from one country. We required the respondents to be residing in the US/UK and had more than 95% task acceptance on both platforms.

We manually removed incomplete or randomly filled responses based on their quality. For instance, some respondents pasted the same text (e.g., a long paragraph about a product, which is completely irrelevant to our topic) to each answer. Others entered unintelligible characters or words. We removed 15 responses from MTurk (used for US respondents) and two responses from Prolific (used for UK respondents) due to quality issues. After filtering, we had 392 valid responses in total, including 192 from the US and 200 from the UK. For our US sample, 61% were male and 39% were female. Respondents' ages ranged from 25 to 35 (SD = 1.15).

88% of the respondents had at least some college education, and 21% had a technology background. For our UK sample, 62% were female and 38% were male. Their ages ranged from 25 to 34 (SD = 1.15). 70% had a college education and 7% had a technology background. The average time for completing the survey was about 25 minutes. After quality check, each US participant was paid $3 and each UK participant was paid $3.5 (to meet the payment requirement of Prolific).

4 Results

This paper focuses on (1) peoples' understanding and misunderstandings of the SADM right; (2) attributes of SADM that our respondents desired to understand; (3) challenges anticipated by the respondents in exercising this right. We found that the UK and US responses were broadly consistent and did not observe any notable differences in these results between the two groups. As detailed in Section 3.3, we conducted tests of proportions to compare the UK vs. US responses and found no statistically significant difference.

4.1 Peoples' Understanding of The Right

4.1.1 Before watching the video tutorial

We asked our respondents about their understanding of the SADM right, before watching the video tutorial of the right. As summarized in Table 1, we present peoples' misunderstandings and reasonable understandings of the right based on the GDPR definition of the SADM right, both before and after watching the video tutorial.

Misunderstandings. Many respondents incorrectly assumed that the SADM right inherently allows people to deny being subjected to SADM or let the companies use their personal information, including profiling for any SADM decision-making. For instance, P148 from UK thought that *"it's the right for you to not consent to automatised decisions"*. Some respondents even anticipated that the right could allow them to completely opt-out of automated decisions made by computers or algorithms. However, this is a misconception because companies can legally do user profiling and SADM if consumers agree to the service contract. In practice, this may allow any Internet-based services (with a valid service agreement and a privacy notice) to subject people to SADM processing. Furthermore, respondents thought that they could choose *ex-ante* (i.e., before a decision is made) between automated and human-involved decision making. For instance, P151 from UK believed that, *"you have the right to request that a human looks at your application for something before a decision is made."* However, the GDPR (under Art 22(3) [70]) allows for human intervention *ex-post* (i.e., only after) the user is subjected to SADM decisions. Majority of respondents reported a lack of understanding of the SADM right, primarily, because they had not heard this right before.

Before tutorial	US	UK
Deny subjecting to SADM	58	57
Human involvement	49	33
Don't know/unsure	45	70
User control and choice	25	25
Prohibit profiling	11	3
Inform about SADM	10	4
Obtain explanation	6	12

After tutorial	US	UK
Deny subjecting to SADM	66	60
Inform about SADM	43	40
Obtain explanation	34	44
Human involvement	27	29
Prohibit profiling	16	11
Don't know/unsure	16	21
Contest against SADM	6	7

Table 1: The top table shows our respondents' answers to the open-ended question, understanding of the SADM right before watching the tutorial. The bottom table shows our respondents' answers to the open-ended question, understanding of the SADM right after watching the tutorial.

Reasonable understandings. Some respondents correctly assumed that they could request for human review of automated decisions ex-post, i.e., after the automated decision is made. Art 22(3) under GDPR provides the right to the consumers to obtain human intervention, express their concern, or even contest the solely-automated decision. We also found people mostly preferred human review in the case of negative outcome (e.g., loan rejection) of a SADM decision. For instance, according to P108 (UK), *"if I have applied for something and have been rejected automatically by [a] computer I can appeal and have it looked at by human."*

Some respondents also preferred to be informed when subjected to SADM and obtain an explanation for SADM decisions made. Art 13(2)(f) GDPR highlights these requirements, stating that companies must inform users about the existence of SADM, including profiling, and provide meaningful information about the logic involved to make the decision.

"I don't know." Many respondents reported having no prior knowledge of the right. While some respondents guessed a basic meaning of the right. Others were unable to even guess. For instance, P36 from US reported that, *'I have not got even the smallest of clues'* about what SADM right means.

4.1.2 After watching the video tutorial.

After watching the tutorial, over 97% of US and 90% of UK respondents correctly answered the attention-checking question, suggesting that the vast majority of them watched the video carefully. They also felt the video was informative in

helping them understand SADM. For example, P85 from US found that *"Some of the questions, such as the "automated decision making" were a little confusing but [..] the video made it a lot easier to understand."* Our results also suggest that most respondents gained a basic understanding of the right after watching the video. Based on the responses from the two multiple-choice questions examining their understanding of the right after watching the video, 80% of both US and UK respondents' expressed understanding was consistent with the legal definition. The enhanced understanding of the SADM right also helped respondents to answer subsequent questions such as those related to attributes of SADM process to explain and challenges to exercise SADM right.

At the end of the survey, over 80% of US respondents and 64% of UK respondents reported either extremely clear or somewhat clear understanding of the SADM right, based on the responses to the Likert scale question. Our respondents also provided their self-reported understanding of the right through an open-ended question towards the end of the survey. We found that their understanding was consistent with the legal definition and covered significant aspects of the right. These aspects included participants' assumptions that the right allows people to be informed about being subjected to SADM, to obtain explanations of the SADM process, to request human involvement, and to contest against solely automated decisions. However, some respondents still had misunderstandings after watching the video. For instance, some of them still believed that the right prohibits user profiling for targeted ads. A large proportion of respondents were still keen to deny being subjected to SADM if it affects them significantly. For instance, P57 from UK reported that his opinion about this right remained same as before watching the video, i.e., *"This is a person's right to get fair treatment free from pre-programmed decision making algorithms."* They prefer not to be subjected to SADM process because they lack trust in the SADM systems to make fair decisions, especially in case of high-stake decisions impacting them legally or significantly.

4.2 SADM Attributes People Want to Know

To capture people's expectations about the attributes of the 'solely' automated decision-making that they desire to understand, we asked an open-ended question, followed by a multiple-choice question. Table 2 summarizes the answers to the two questions, respectively. These answers are attributes frequently mentioned by our respondents.

4.2.1 Attributes from survey open-ended responses

First, we present the results from the open-ended question. We further classify them as part of the local or global explanation (as defined in section 2.3), wherever applicable. Our respondents' answers (summarized in top of Table 2) covered major themes such as the type of information used, the process of

Open-ended	US	UK
Type of information	77	88
Personalized explanation	56	52
Unique factors	44	34
Source of information	35	56
No explanation	12	20
Human involvement & appeal	19	24
Fairness of algorithm	6	10

Multiple-choice	US	UK
Type of information	40	36
Explaining logic involved	34	36
Fairness of algorithm	34	35
Source of information	34	31
Company infrastructure	32	29
Data accuracy and updates & appeal	28	27

Table 2: The top table shows our respondents' answers to the open-ended question regarding what attributes they would like to understand and the corresponding number of US and UK respondents who mentioned each attribute. The bottom table shows our respondents' answers to the multiple-choice question where they can select multiple answers where each answer option (i.e., attribute to explain) was suggested by the policy-makers (in this case, the Working Party A29WP [26]). In the top table, the attributes in bold were only expressed by the respondents, i.e., not mentioned by A29WP and thus not shown in the right table.

SADM, and the source of information. A large percentage of respondents (US 40%, UK 44%) reported that they were interested in knowing about the 'type of information' used. They also desired to know the 'source of information' (US 18%, UK 28%). For instance, P136 from UK wanted to *"[..] know what information is used to make the decision and where it was obtained from."*

Furthermore, many respondents desired to know the 'unique factors considered' (US 23%, UK 17%), i.e., the criteria used in the decision-making and the *weights* of those factors that were unique to their profiles. For instance, P46 from the US emphasized that *"I would like for them to explain how much weight they put on what specific data information that have about me"* and *"how they determine what data is important and what data is not."* Similarly, P14 from UK requested to know *"what factors are taken into consideration when applying for a loan ...[and what] could hinder your chances of getting a loan."* 'Factors considered' differ from the 'type of information' used because the factors could include a subset of the user information collected and possibly other non-user information (e.g., market interest rate) as criteria for decision-making. For instance, a company can collect

different types of user information such as age and gender, but age is the most important factor considered for SADM (e.g., age greater than a certain number). These attributes can be classified as part of local explanation since they help to explain how the model makes a specific decision.

The need to know the factors considered relates to the expectations to receive *personalized* explanations. People reported that they wanted to have personalized explanations of automated decisions (US 29%, UK 26%) as a way to seek justification for the decision made for them. This explanation should include reasons for a negative decision, their impact, limitations, and ways to improve/alter the decision. For instance, P136 from UK wanted to know *"How my circumstances were assessed - what factors led to me being accepted or turned down, if it was a points based system, how close was I/ what would I need to do to achieve the necessary score."* She preferred a personalized explanation of a particular decision rather than a general explanation of the SADM process. She also desired actionable information about what she can do to improve the decision outcome. Similarly, P170 from US stated that *"It would be the same questions I would ask of an employee"*. Here P170 expected to receive a machine-provided explanation similar to the experience of receiving an explanation from a human decision-maker. One possibility is to create chatbots that can converse with the user to explain the decision made automatically. In another instance, both P51 from US and P195 from UK pointed out the risks of user profiling in SADM and therefore expected greater transparency in explanations. P51 suspected that the automated decisions could use *"the data that was biased in any way in an unfair manner"* and therefore demanded to know *"complete description of how each aspect causes a change in their decisions [..] as well as the changes that could be made"*. Similarly, P195 questioned *"how certain features such as age and gender can"* influence a decision to make *"people be potentially deprived a loan"*.

It is worth noting that some respondents expressed that they did not want any explanation about automated decision-making (US: 12, UK: 20). They did not report specific reasons for this preference when answering this question. However, later when discussing their anticipated challenges in exercising the right, some respondents said that knowing SADM can be boring, or not worth their time and effort. For instance, P145 from the UK said, *"effort required to exercise this right is greater than the 'reward' I would expect to gain."* Here, he seemed to derive his preference based on a cost-benefit analysis of exercising the right. As we will discuss later, one common challenge people anticipated was that exercising the right can be too time-consuming. Another possible reason could be that people feel resigned about SADM. Privacy scholars [67] have suggested that most Americans give up data for relevant ads, not because of convenience, but resignation. Rather than participating in a rational exchange, consumers are giving up their personal information with 'a feeling of futility' [67]. Future research can further investigate why some people do not want an explanation of SADM.

4.2.2 Attributes in policy-makers' recommendations

Next, we compare themes from the open-ended question with the six pre-defined attributes from the MCQ-based question. As described in section 3.2, these pre-defined attributes were adapted from the GDPR Working Party-29 Guidelines [26] and included: a) type of information, b) explaining logic involved, c) fairness of algorithm, d) source of information, e) company infrastructure, and f) data accuracy and updates.

Table 2 (right) shows US and UK respondents' choices of these pre-defined attributes of SADM where each respondent can choose multiple attributes. A majority of respondents from both countries wanted to know the 'type of information' used, followed by 'explaining logic involved', 'fairness of algorithm', and 'source of information.' 'Data accuracy and updates' and 'company infrastructure' were least selected.

It is interesting to note that three major themes from the open-ended question corroborate with similar pre-defined attributes from the multiple-choice question in terms of how commonly they were expected by our respondents. These included 'type of information', 'explaining logic involved' and the 'source of information' used.

However, we also observed some differences between the responses to the two questions. For instance, while about one-sixth of respondents selected 'Fairness of algorithm' in the multiple-choice question, a much smaller percentage of respondents expressed it in the open-ended question. Since we asked the open-ended question first, this might suggest respondents did not immediately think about the fairness aspect. Or in other words, they might be paying more attention to a negative decision about themselves and a need for a personalized explanation of the decision rather than whether the SADM process is fair.

Last but not least, we found that some themes were unique to the open-ended question, such as 'seeking justification' for the decision, requesting a personalized explanation including the weight of the 'factors considered', or not willing to be presented with any explanation at all. This is an important list of attributes because they were sought by the respondents but were not covered in policy makers' recommendations. This gap might lead to ineffective SADM explanations that do not satisfy people's expectations or needs.

4.3 Anticipated Challenges

We heard from our respondents about what attributes they prefer in the explanations for the SADM decision. However, will they be able to exercise this right easily if they want to? We next asked respondents about their perceived challenges of using this right. Participants reported several types of challenges that may arise at different stages of exercising the

Major challenges anticipated	US	UK
Hard to safeguard against SADM processing	40	54
Common Challenges	63	75
Hard to identify and fight bias in algorithm	62	69
Hard to contest SADM decisions	32	29
Review process not user friendly	26	18

Table 3: Main perceived challenges of exercising SADM mentioned by our respondents.

SADM right. Some of these challenges (e.g., hard to rectify incorrect information about individuals) have been reported in other contexts. In contrast, other challenges seem more salient in the context of the SADM right. We will briefly summarize the former and then focus on the latter. Table 3 summarizes the number of US and UK respondents who mentioned each type of these challenges.

4.3.1 Common Challenges

Our respondents reported several perceived challenges of the SADM right. These challenges are also common in other contexts for protecting people's privacy, such as: lack of user awareness as well as control of personal data collection and sharing, lack of user trust on companies, lack of transparency of company privacy practices, time-consuming to communicate with companies, and difficult to rectify incorrect information that companies have about individuals.

4.3.2 Challenges More Salient in SADM

From the participants' responses, we identified several challenges that are either more salient or have different implications in the context of SADM right. We present them below.

Hard to identify and fight biases in algorithms. Respondents were concerned that the algorithms used in the SADM process could be wrought with biases. Such a bias could be hard to detect, verify, prove to cause discrimination. P184 from the US pointed out that as an initial step, it will be difficult to check *"how the algorithm is updated and how fair and unbiased it is."* He further suspected that the companies could be reluctant to provide such information, making it all the more difficult to assess. Furthermore, if the algorithm is found to be biased, it may be challenging to provide proof of bias to the company and request them to make it unbiased for future decision-making.

Hard to safeguard against SADM processing. Many participants reported that they might not be aware of being subjected to SADM processing while using an internet-based service, e.g., social media, banking. For instance, P 136 from the US expressed that *"It might be difficult to know in the first place whether I was affected by automated decision making or not."* It could be because companies might not actively notify people about SADM processing or, as P21 from the US noted, *"It's hidden in legal terms that most people don't understand so you don't ever even notice it."* Additionally, respondents felt that once the SADM system decides for them, it would be challenging to request a re-evaluation of the outcome. Users might find it challenging to explain their circumstances for companies to re-evaluate the decisions about them. For instance, P36 from the UK explained that while applying for a job or a loan, one of the biggest challenges is that SADM *"decisions [..] don't take into consideration personal circumstances or personality"* of an individual and that the person is *"not being able to explain [themselves].."* in case of negative outcome. As a result, people expected to have an ex-ante opt-out option to not be subjected to the SADM processing. However, they also worried that even if the option existed, they may not know it either due to their own blind-spot or the lack of transparency from companies.

Hard to contest SADM decisions. Respondents from both the UK and the US pointed out that it will be difficult to contest the SADM decisions. Companies can show resistance or be *"hesitant about having human review the decision of the automated system"* as P105 (US) put it. He anticipated that one way of doing so is to make it difficult to contact the company for human involvement in SADM. Based on his real life experiences, he cited that *"unfortunately, most customer service is anonymously automated. Seldom is there an actual human being in which I could contact to deal with this issue."* While reaching customer services might be a need and challenge in commerce, getting a real person to check the machine-made decisions is at the core of the SADM right.

Review process not user-friendly. Even if the users get access to the customer representative, such customer staff may lack relevant training to review the SADM decisions. For instance, P81 from US questioned the quality of human involvement to review a decision. He expressed, *" I don't anticipate any meaningful human interventions that would contradict the results of an automated system."* It could also be possible that companies may not invest in a dedicated team of experts to review automated decision, and *"presumably, the "human intervention" could just be an intern who clicks okay to a computer prompt"*. Additionally, the companies may even lack human resources to respond to large amounts of complaints, leading to slow redress of decision.

People were also concerned that even with required human involvement, the review process *"won't be impartial"* (P80, UK). P88 from the US suspected companies would be *"unwilling to change decision"* because *"[it] likely makes them too much money for them to actually be willing to do anything to change it."* This highlights people's lack of trust in companies to carefully review SADM decisions and change them if needed.

5 Discussion

This study explored research questions regarding people's understanding of the SADM right, aspects of SADM that they want to be explained, and their perceived challenges in exercising the SADM right. Our respondents incorrectly assumed that the right allows them to deny being subjected to SADM or to opt-out altogether. Respondents expected a few novel attributes of SADM to be explained, which are not covered by government guidelines on SADM [26]. These novel attributes ranged from receiving personalized explanations to seeking justification for adverse outcomes. Lastly, our respondents reported four broad perceived challenges of exercising the SADM right: i) *Hard to safeguard against SADM processing*: respondents anticipated that they might not be aware of whether or when companies subject them to SADM processing; ii) *Hard to identify and fight algorithmic biases*: respondents were concerned that algorithms might harbor biases that would be hard to detect, verify, and prove to cause discrimination; iii) *Hard to contest the outcome of SADM*: respondents anticipated that companies could show resistance to review adverse outcomes/decisions by complicating ways to contact them or denying human reviewers to re-evaluate the SADM decision; and iv) *Unfriendly review process*: respondents anticipated that companies may lack the infrastructure to respond to SADM decision review requests or that human reviewers may lack the relevant training to conduct the review. Table 4 summarizes the main findings. Next, we will discuss design and policy implications that can mitigate some of the misunderstandings and challenges and support people's informational needs of SADM.

People's perspective on SADM right	Main findings
Understandings of the SADM right	**Misunderstandings:** - Deny subjecting to SADM - User Choice and Control (e.g. opt out) - Don't know **Reasonable understandings:** - Obtain explanation - To be informed about decision - Prohibit profiling (e.g. for targeted ads) - Request for human intervention - Contest against SADM right
Attributes of SADM that people desired to be explained	**Attributes (open-ended):** - Type of information - Personalized explanation - Unique factors - Source of information - Human involvement and appeal - Fairness of algorithm
Challenges anticipated in exercising the SADM right	- Hard to safeguard against SADM processing - Hard to identify and fight bias in algorithm - Hard to contest SADM decisions - Review process not user friendly - Other common challenges

Table 4: A summary of the main findings from the study.

5.1 Design Implications

5.1.1 Help People Understand SADM

Personalized explanation templates. Our findings suggest that people have different expectations or informational needs for explanations of SADM. Many respondents anticipated personalized explanations of the SADM outcomes. They expect companies to provide explanations that can justify the decision regardless of the outcome and provide actionable suggestions on what people may do to improve the decision outcome, especially in negative decisions. While designing explanations, caution must be taken to balance over-simplified vs. over-complicated explanations [25, 40, 49]. To strike a good balance, companies can present hierarchical explanations with multiple levels of details and personalization, which is similar to the privacy nutrition labels for IoT devices [23]. For instance, it would be useful to design explanation templates that would allow people to zoom into details about a specific decision, personalized based on individuals' demographics (e.g., age) and computing knowledge. It could also include suggestions to improve the outcome next time. Suppose a bank rejects a person for a credit card; it could provide an explanation template with multiple levels of personalized explanations. At the basic level, it could include the outcome and a high-level reason (e.g., low credit score). At the intermediate level, the user could zoom into details such as attributes considered for the outcome and which attributes the user could improve for a better outcome. A more detailed level could include additional information such as what other personal data the company uses to make SADM decisions and its source, how the company ensures the fairness and accuracy of the algorithm used, and other aspects (see Table 2).

Sandboxes to play with SADM systems. Another key finding of our study is that our respondents expected explanations to include how a decision is relevant to them. It includes 1) describing the (significant) effect that the SADM decision has on the individual and 2) providing individually tailored and practically actionable recommendations to improve future outcomes for the individual. A concrete design idea is that companies can implement interactive interfaces that describe how different factors affect a decision, allow users to interact with various factors to see their impact and provide personalized recommendations to improve acceptance for future outcomes. For instance, a car insurance company can provide customers with an interactive sandbox to explain the extra insurance premium for certain events (e.g., speeding). It can also allow customers to play with the algorithm (or sandbox) by testing different factors such as levels and times of speeding, driver age, and past accident records to see the dependence of outcome on these factors. However, as a challenge, the sandbox could possibly allow for reverse-

engineering the model, which would not be ideal in anomaly detection cases (e.g., fraud prevention).

5.1.2 Mitigate Misunderstandings and Challenges

Icons for SADM processing. Since our respondents had many misconceptions about the SADM right, it would be beneficial for companies to consider these misconceptions when designing their platforms to better support users exercise the SADM right. Additionally, our respondents anticipated that it would be challenging to detect whether/when they are subject to SADM. One way to communicate whether a user is subject to SADM is to show indicators of SADM (visual icons or other modalities). Various privacy icons have been proposed to convey complicated privacy concepts. Some of the icons are about the privacy notices in various domains (e.g., online tracking [19,47], social media [36], web cams [22,58], web links [37]), while others are to convey privacy choices [33,65].

It is worth noting that even though there are icons representing targeted ads, targeted ads are only one example of a much broader set of SADM practices. We are not aware of any existing icons for SADM processing. For instance, imagine that a bank website shows a SADM icon next to its credit card application and loan application. An e-commerce site shows the icon next to its recommended products to a user, or a social media site displays the icon next to the recommended friends to a user. These icons can represent whether the corresponding decisions about the prevailing user are made by an automated system using algorithms alone (SADM) or by involving humans in the process (human-AI hybrid).

In addition, these icons could increase transparency by highlighting the uncertainty of the decision made by using solely automated systems. This was another aspect that some of our respondents reported as a challenge of the right. The icons would be even more useful if they are standardized (e.g., by self-regulating trade organizations or the Internet standard organization, W3C). While these icons are likely to have their challenges (e.g., people may ignore or misunderstand them), when they are designed and evaluated appropriately (e.g., see good examples of nutrition labels and the recent CCPA opt-out icons), they can help communicate SADM.

Social support for contesting algorithmic decisions. Our respondents felt it would be challenging to contest the outcome of SADM due to unfriendly review processes. They also anticipated the difficulty for them to explain their personal circumstances to appeal a negative decision. One design direction is to create tools or platforms that allow people to share and learn from each other about their strategies and experiences in working with specific companies for their redress/contest requests. This is similar to crowd-sourcing help for tech support, such as [46]. End-user tools could be designed to guide users to create and share contest requests (e.g., answering a set of questions and attaching supporting documents). These shared user experiences can also further motivate companies to improve how they handle people's redress requests. How SADM systems can be designed to support contestability is another exciting future direction.

5.2 Policy Implications

Our results also have some policy implications. For instance, policymakers could consider adding novel attributes to policy guidelines for companies to explain decisions made by SADM. As suggested by participants, these attributes include 'seeking justification' for the decision, requesting 'personalized explanation' including the weight of the 'factors considered,' or an option to opt-out of receiving any explanation at all. Policymakers could also consider mandating notifications to alert users when subjected to SADM processing, similar to website cookie notifications. Lastly, policymakers and industry organizations could initiate regulatory standards that could be applied internationally (e.g., a standard explanation template) similar to ISO standards.

6 Limitations

Our study has several limitations. First, we recruited adult participants, but our participants ended up with a small age range (about 25-35 years old), which was not intentional. Second, our study data does not explain the lack of differences between the US and UK responses. Third, we analyzed but did not observe any correlations between the "no explanation" responses and other participant data (e.g., demographics). Fourth, our results may not be generalizable since we cannot claim that our sample represents the populations in those two countries. Lastly, our study focused on people's desire for explanation rather than how SADM explanations are currently implemented. Future work could attempt to address these limitations and explore these directions further.

7 Conclusion

Automated decision-making, including profiling based solely on computer algorithms, continues to grow in importance as more companies adopt this practice to improve efficiency. The challenge, therefore, is to find the balance between opportunities for the companies and the impact on end-users. We studied people's understandings and expectations of the right against solely automated decisions because they are major stakeholders and would be directly impacted by these automated decisions. We presented design implications for companies to support citizens in exercising this right. Future research can explore concrete designs and modalities to explain various attributes of automated decision-making that people desire to understand.

8 Acknowledgement

We thank our participants for sharing their insights. We also thank the anonymous reviewers for their thoughtful feedback.

References

[1] 2018. The EU General Data Protection Regulation | Human Rights Watch. (June 2018). (Accessed on 09/12/2018) https://www.hrw.org/news/2018/06/06/eu-general-data-protection-regulation.

[2] Mike Ananny and Kate Crawford. 2018. Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society* 20, 3 (2018), 973–989.

[3] Yannis-NYU Bakos, David R Trossen, and others. 2009. Does Anyone Read the Fine Print? Testing a Law and Economics Approach toStandard Form Contracts. (2009).

[4] Solon Barocas. 2014. Data mining and the discourse on discrimination. In *Data Ethics Workshop, Conference on Knowledge Discovery and Data Mining*. 1–4.

[5] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 377.

[6] Rainer Böhme and Stefan Köpsell. 2010. Trained to accept?: a field experiment on consent dialogs. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2403–2406.

[7] Mark Bovens. 2007. Analysing and assessing accountability: A conceptual framework 1. *European law journal* 13, 4 (2007), 447–468.

[8] Richard E Boyatzis. 1998. *Transforming qualitative information: Thematic analysis and code development*. sage.

[9] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *Healthcare Informatics (ICHI), 2015 International Conference on*. IEEE, 160–169.

[10] Corinne Cath, Sandra Wachter, Brent Mittelstadt, Mariarosaria Taddeo, and Luciano Floridi. 2018. Artificial intelligence and the 'good society': the US, EU, and UK approach. *Science and engineering ethics* 24, 2 (2018), 505–528.

[11] Electronic Privacy Information Center. 2016. EPIC - AI and Human Rights. (2016). https://www.epic.org/ai/index.html (Accessed on 09/17/2018).

[12] Robert P Charrow and Veda R Charrow. 1979. Making legal language understandable: A psycholinguistic study of jury instructions. *Columbia law review* 79, 7 (1979), 1306–1374.

[13] Danielle Keats Citron and Frank Pasquale. 2014. The scored society: Due process for automated predictions. *Washington Law Review* 89 (2014), 1.

[14] Cary Coglianese and David Lehr. 2016. Regulating by robot: Administrative decision making in the machine-learning era. *Georgetown Law Journal* 105 (2016), 1147.

[15] European Commission. 2018. Can I be subject to automated individual decision-making, including profiling? https://ec.europa.eu/info/law/law-topic/data-protection/reform/rights-citizens/my-rights. (August 2018).

[16] Thomas H Davenport and Jeanne G Harris. 2005. Automated decision making comes of age. *MIT Sloan Management Review* 46, 4 (2005), 83.

[17] Culture Media Sports Department for Digital. 2017. A New Data Protection Bill: Our Planned Reforms. (August 2017). (Accessed on 09/16/2018) https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/635900/2017-08-07_DP_Bill_-_Statement_of_Intent.pdf.

[18] Culture Media Sports Department for Digital. 2019. The Data Protection, Privacy and Electronic Communications (Amendments etc) (EU Exit) Regulations 2019. (February 2019). (Accessed on 02/23/2021) https://www.legislation.gov.uk/uksi/2019/419/made.

[19] Disconnect, Inc. 2014. Disconnect Privacy Icons. (2014). https://github.com/disconnectme/privacy-icons.

[20] Pam Dixon. 2008. A Brief Introduction to Fair Information Practices. (2008). https://tinyurl.com/7mmym54v (Accessed on 09/17/2018).

[21] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. 2017. Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134* (2017).

[22] Serge Egelman, Raghudeep Kannavara, and Richard Chow. 2015. Is This Thing On? Crowdsourcing Privacy Indicators for Ubiquitous Sensing Platforms. In *Conference on Human Factors in Computing Systems (CHI)*. ACM, 1669–1678. DOI:http://dx.doi.org/10.1145/2702123.2702251

[23] Pardis Emami-Naeini, Yuvraj Agarwal, Lorrie Faith Cranor, and Hanan Hibshi. 2020. Ask the Experts: What Should Be on an IoT Privacy and Security Label?. In *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE, 447–464.

[24] Motahhare Eslami, Sneha R Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. 2018. Communicating Algorithmic Process in Online Behavioral Advertising. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 432.

[25] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. " Be Careful; Things Can Be Worse than They Appear": Understanding Biased Algorithms and Users' Behavior Around Them in Rating Platforms.. In *ICWSM*. 62–71.

[26] Working Party EU. 2018. Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679. (February 2018). (Accessed on 09/17/2018) https://iapp.org/media/pdf/resource_center/W29-auto-decision_profiling_02-2018.pdf.

[27] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical methods for rates and proportions*. John Wiley & Sons.

[28] Tarleton Gillespie, Pablo J Boczkowski, and Kirsten A Foot. 2014. *Media technologies: Essays on communication, materiality, and society*. MIT Press.

[29] Bryce Goodman and Seth Flaxman. 2017. European Union regulations on algorithmic decision-making and a" right to explanation". *AI magazine* 38, 3 (2017), 50–57.

[30] Robert Gorwa. 2019. What is platform governance? *Information, Communication & Society* 22, 6 (2019), 854–871.

[31] Nina Grgic-Hlaca, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Lyon, France, 903–912. DOI:http://dx.doi.org/10.1145/3178876.3186138

[32] Jens Grossklags and Nathan Good. 2007. Empirical studies on software notices to inform policy makers and usability designers. In *International Conference on Financial Cryptography and Data Security*. Springer, 341–355.

[33] Hana Habib, Yixin Zou, Yaxing Yao, Alessandro Acquisti, Lorrie Cranor, Joel Reidenberg, Norman Sadeh, and Florian Schaub. 2021. Toggles, dollar signs, and triangles: How to (in) effectively convey privacy choices with icons and link texts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–25.

[34] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, Miro Dudik, and Hanna Wallach. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, Glasgow, Scotland Uk, 1–16. DOI:http://dx.doi.org/10.1145/3290605.3300830

[35] Xia Mengnan Du Hu, Ninghao Liu. 2020. Techniques for Interpretable Machine Learning. *Commun. ACM* 63, 1 (2020), 68–77. https://cacm.acm.org/magazines/2020/1/241703-techniques-for-interpretable-\machine-learning/fulltext

[36] Renato Iannella and Adam Finden. 2010. Privacy Awareness: Icons and Expression for Social Networks. In *International Workshop for Technical, Economic and Legal Aspects of Business Models for Virtual Goods*. Article 1, 15 pages. http://virtualgoods.org/2010/VirtualGoodsBook2010_13.pdf.

[37] Saraschandra Karanam, Janhavi Viswanathan, Anand Theertha, Bipin Indurkhya, and Herre Van Oostendorp. 2010. Impact of Placing Icons Next to Hyperlinks on Information-Retrieval Tasks on the Web. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 32. eScholarship, 2834–2839. https://escholarship.org/content/qt27w0n9kc/qt27w0n9kc.pdf.

[38] Farzaneh Karegar. 2018. *Towards Improving Transparency, Intervenability, and Consent in HCI*. Ph.D. Dissertation. Karlstad University Press.

[39] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 453–456. DOI:http://dx.doi.org/10.1145/1357054.1357127

[40] René F Kizilcec. 2016. How much information?: Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2390–2395.

[41] Joshua A Kroll, Solon Barocas, Edward W Felten, Joel R Reidenberg, David G Robinson, and Harlan Yu. 2016. Accountable algorithms. *U. Pa. L. Rev.* 165 (2016), 633.

[42] Min Kyung Lee and Su Baykal. 2017. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, Portland, Oregon, USA, 1035–1048. DOI: http://dx.doi.org/10.1145/2998181.2998230

[43] Tae Ho Lee and Lois A Boynton. 2017. Conceptualizing transparency: Propositions for the integration of situational factors and stakeholders' perspectives. *Public Relations Inquiry* 6, 3 (2017), 233–251.

[44] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, David Madigan, and others. 2015. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics* 9, 3 (2015), 1350–1371.

[45] David Meyer. 2019. Who should enforce a US federal privacy law? https://iapp.org/news/a/who-should-enforce-a-federal-privacy-law/. (2019). (Accessed on 02/23/2021).

[46] Microsoft. 2019. Microsoft Community. https://answers.microsoft.com/en-us/page/gettingstarted. (2019). (Accessed on 02/23/2021).

[47] Mozilla. 2020. Privacy Icons. (February 2020). https://wiki.mozilla.org/Privacy_Icons.

[48] Deirdre K Mulligan, Joshua A Kroll, Nitin Kohli, and Richmond Y Wong. 2019. This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–36.

[49] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682* (2018).

[50] Commission nationale de l'informatique et des libertes. 2013. Homepage | CNIL. (2013). (Accessed on 09/17/2018) https://www.cnil.fr/en/home.

[51] Daniel Neyland. 2016. Bearing account-able witness to the ethical algorithmic system. *Science, Technology, & Human Values* 41, 1 (2016), 50–76.

[52] International Association of Privacy Professionals (IAPP). 2020. The California Privacy Rights Act of 2020. https://iapp.org/resources/article/the-california-privacy-rights-act-of-2020/. (2020). (Accessed on 02/23/2021).

[53] Information Commissioner's Office. 2018. Home | ICO. (2018). (Accessed on 09/17/2018) https://ico.org.uk/.

[54] Stefan Palan and Christian Schitter. 2018. Prolific.ac - A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.

[55] Andrew S Patrick and Steve Kenny. 2003. From privacy legislation to interface design: Implementing information privacy in human-computer interactions. In *International Workshop on Privacy Enhancing Technologies*. Springer, 107–124.

[56] Sam Pfeifle. 2018. California passes landmark privacy legislation. https://tinyurl.com/b5es9y9x. (2018). (Accessed on 02/23/2021).

[57] Irene Pollach. 2007. What's wrong with online privacy policies? *Commun. ACM* 50, 9 (2007), 103–108.

[58] Rebecca S Portnoff, Linda N Lee, Serge Egelman, Pratyush Mishra, Derek Leung, and David Wagner. 2015. Somebody's Watching me? Assessing the Effectiveness of Webcam Indicator Lights. In *Conference on Human Factors in Computing Systems (CHI)*. ACM, 1649–1658. DOI:http://dx.doi.org/10.1145/2702123.2702164

[59] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1135–1144.

[60] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry* (2014), 1–23.

[61] Nripsuta Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David Parkes, and Yang Liu. 2019. How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. http://arxiv.org/abs/1811.03654 arXiv: 1811.03654.

[62] Nick Seaver. 2013. Knowing algorithms. *Media in Transition* 8 (2013), 1–12.

[63] Andrew Selbst and Solon Barocas. 2017. Regulating inscrutable systems. *WeRobot 2017* (2017).

[64] Donghee Shin and Yong Jin Park. 2019. Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior* 98 (2019), 277–284.

[65] Stanford Legal Design Lab. 2020. Icons for legal help. (2020). https://betterinternet.law.stanford.edu/design-guide/icons-for-legal-help/.

[66] Alan B Tickle, Robert Andrews, Mostefa Golea, and Joachim Diederich. 1998. The truth will come to light: Directions and challenges in extracting the knowledge embedded within trained artificial neural networks. *IEEE Transactions on Neural Networks* 9, 6 (1998), 1057–1068.

[67] Joseph Turow, Michael Hennessy, and Nora Draper. 2015. The tradeoff fallacy: How marketers are misrepresenting American consumers and opening them up to exploitation. *Available at SSRN 2820060* (2015).

[68] UK Information Commissioner's Office (ICO). 2019. *ICO and The Turing consultation on Explaining AI decisions guidance*. Technical Report. https://tinyurl.com/2dk7rvrd Publisher: ICO, (Accessed on 02/23/2021).

[69] UK Information Commissioner's Office (ICO). 2020. *Explaining decisions made with AI*. Technical Report. https://tinyurl.com/ico-explaining-decisions Publisher: ICO, (Accessed on 02/23/2021).

[70] European Union. 2016. Regulation 2016/679 of the European parliament and the Council of the European Union. *Official Journal of the European Communities* 2014, March 2014 (2016), 1–88. DOI:http://dx.doi.org/pri/en/oj/dat/2003/l_285/l_28520031101en00330037.pdf

[71] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, Montreal QC, Canada, 1–14. DOI:http://dx.doi.org/10.1145/3173574.3174014

[72] Maranke Wieringa. 2020. What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 1–18.

[73] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of CHI2018 (CHI '18)*. Association for Computing Machinery, Montreal QC, Canada, 1–14. DOI:http://dx.doi.org/10.1145/3173574.3174230

Survey Questions:

Q1 Have you heard about the “Right against solely automated decision making" before?

- ○ Yes (1)
- ○ No (2)

Q2 What do you think the “Right against solely automated decision making" means? Pleasebriefly describe your understanding below.

_____

_____

_____

Q3 Please: 1) Briefly describe your understanding of "User Profiling"?
2) Give an example to illustrate your understanding.

_____

_____

_____

Q4 Please: 1) Briefly describe your understanding of "Automated decision making"?
2) Give an example to illustrate your understanding.

_____

_____

_____

Q5 Now we will explain the meaning of the ***“Right against solely automated decision making***". Please either read the description or watch the video carefully. The next few questions will be based on your understanding from the resources below.

Video: youtu.be/BrQMqmPEWQs

Q3.6 In the video, did you spot a computer screen?

- ○ Yes (1)
- ○ No (2)

Q7 Do you have any suggestions to improve this video? Please write them down below.

_____

_____

_____

Q8 Based on the information you have just learned from the video/text description, which of the following statements best explains the “Right against solely automated decision making"? (Choose all that apply)

- ☐ It allows users to request companies to delete their personal data (1)
- ☐ It allows users to request companies to explain how the AI makes automated decisions about them (2)
- ☐ It allows users to object to the processing of their personal data (3)
- ☐ It allows users to request companies to correct their personal data (4)
- ☐ Other (Please Specify) (5) _____

Q9 Based on the information you have just learned from the video/text description, which of the following is an example of the “Right against solely automated decision making"? (choose allthat apply)

- ☐ A college level football player stumbled upon a photo of them on a webpage of a sports magazine. This person wonders how the magazine editor found the photo. (1)
- ☐ An individual stumbled upon the social media settings page that shows an automatically generated profile that the company uses to provide a personalized news feed and other targeted ads. This person requests the company to explain the process of automatic profiling. (2)
- ☐ An individual request the bank to transmit details of this person’s bank transactions to a new Budget planning app. (3)
- ☐ An individual recently switched jobs but still appears on the social events webpage of the previous organization. This person requests the organization to remove such information. (4)

☐ Other (Please specify) (5) _____

Q10
If the companies start to make decisions about you automatically, what kind of decisions do you think might significantly affect you? Please explain briefly.

_____

_____

_____

Q11 Regarding the example from the video about automatically making decisions on bankloans, what aspects of automated decision making would you like to be explained?

_____

_____

_____

Q12 Which of the following aspects of automated decision making would you like to be explained/provided with? (select all that apply)

☐ Company must inform me where or how the company found/obtained my data that is used in making automatic decisions about me (1)

☐ Company must inform me what types of data (e.g., my name, age, address) are used in making automatic decisions about me (2)

☐ Company must inform me how data is used or what algorithm or method is used in making automatic decisions about me (3)

☐ Company must inform me how it ensures that the algorithm or method used in making automatic decisions about me is fair and unbiased (4)

☐ Company must update both my data and the algorithm for accurate decision making as well as inform me about these updates (5)

☐ Company must provide me with customer care services to contest automated decision making (6)

☐ Other, please specify (7) _____

Q13 In your past internet usage experience, have you ever encountered an incident where you could have exercised your “Right against solely automated decision making"?

○ Yes (1)

○ No (2)

Display This Question:

If In your past internet usage experience, have you ever encountered an incident where you could hav... = Yes

Q14 Could you briefly explain this past incident? How could you have exercised this right?

_____

_____

_____

Display This Question:

If In your past internet usage experience, have you ever encountered an incident where you could hav... = Yes

Q15 What are some challenges you think you could have encountered while exercising your “Right against solely automated decision making" for the incident you mentioned in the previous question? Please explain briefly.

_____

_____

_____

Display This Question:

If In your past internet usage experience, have you ever encountered an incident where you could hav... = Yes

Q16 What could be some possible solutions you would suggest to companies in order to overcome the challenges you mentioned in the previous question? Please explain briefly.

_____

_____

Display This Question:

If In your past internet usage experience, have you ever encountered an incident where you could hav... = No

Q17 Based on your understanding of the “Right against solely automated decision making", how do you expect to be benefited from exercising this right online? Please briefly describe below.

_____

_____

_____

Display This Question:

If In your past internet usage experience, have you ever encountered an incident where you could hav... = No

Q18 Based on your understanding of the “Right against solely automated decision making", what are some of the challenges you anticipate while exercising this right? Please briefly describe below.

_____

_____

_____

Display This Question:

If In your past internet usage experience, have you ever encountered an incident where you could hav... = No

Q19 What are some possible solutions you would suggest to companies in order to overcome the challenges you mentioned in the previous question? Please briefly describe below.

_____

_____

_____

Q20 On the scale of 1 to 5, how well were you able to understand the "Right against solely automated decision making"?

	Extremely unclear (25)	Somewhat unclear (26)	Neither clear nor unclear (27)	Somewhat clear (28)	Extremely clear (29)
Understanding of "Right against solely automated decision making" (1)	o	o	o	o	o

Q21 Please briefly describe below your overall understanding of the “Right against solely automated decision making"?

A Fait Accompli? An Empirical Study into the Absence of Consent to Third-Party Tracking in Android Apps

Konrad Kollnig, Reuben Binns, Pierre Dewitte*, Max Van Kleek,
Ge Wang, Daniel Omeiza, Helena Webb, Nigel Shadbolt
Department of Computer Science, University of Oxford, UK
**Centre for IT and IP Law, KU Leuven, Belgium*
firstname.lastname@(cs.ox.ac.uk | kuleuven.be)

Abstract

Third-party tracking allows companies to collect users' behavioural data and track their activity across digital devices. This can put deep insights into users' private lives into the hands of strangers, and often happens without users' awareness or explicit consent. EU and UK data protection law, however, requires consent, both 1) to access and store information on users' devices and 2) to legitimate the processing of personal data as part of third-party tracking, as we analyse in this paper.

This paper further investigates whether and to what extent consent is implemented in mobile apps. First, we analyse a representative sample of apps from the Google Play Store. We find that most apps engage in third-party tracking, but few obtained consent before doing so, indicating potentially widespread violations of EU and UK privacy law. Second, we examine the most common third-party tracking libraries in detail. While most acknowledge that they rely on app developers to obtain consent on their behalf, they typically fail to put in place robust measures to ensure this: disclosure of consent requirements is limited; default consent implementations are lacking; and compliance guidance is difficult to find, hard to read, and poorly maintained.

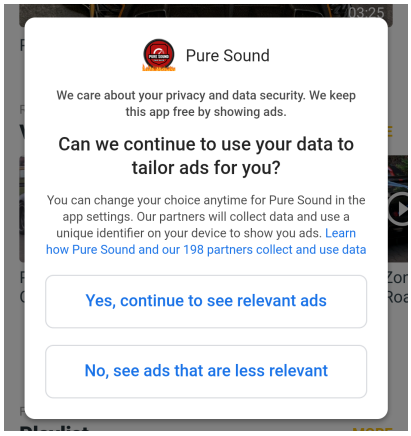
1 Introduction

Third-party tracking, the deliberate collection, processing and sharing of users' behavioural data with third-party companies, has become widespread across both mobile app ecosystems [16, 83, 85] and the web [16, 67]. The use of third-party trackers benefits app developers in several ways, notably by providing analytics to improve user retention, and by enabling the placement of personalised advertising within apps, which often translates into a vital source of revenue for them [32, 62]. However, it also makes app developers dependent on privacy-invasive data practices that involve the processing of large amounts of personal data [40, 48, 62], with little awareness from users and app developers [28, 71, 74, 85]. Data protection and privacy legislation such as the General Data Protection Regulation (GDPR) [38] in the EU and the UK, and the Children's Online Privacy Protection Act (COPPA) [79] in the US, establish clear rules when it comes to the processing of personal data and provide additional safeguards when it comes to information relating to children. As explained in Section 3, consent is a necessary precondition for third-party tracking.

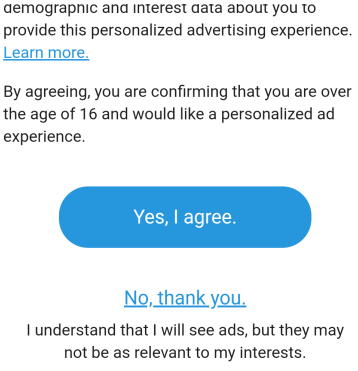
The implementation of consent in mobile apps has—since the end of 2020—sparked a fierce public battle between Apple and Facebook over tracking controls in iOS 14.5 [1, 2]. To give users more control over their data, Apple has introduced an opt-in mechanism for the use of the Advertising Identifier (AdID)—similar to how apps currently request location or contacts access. Facebook, like many other mobile advertising companies, is concerned that most users will not agree to tracking if asked more clearly and explicitly [3]; iOS users could already opt-out from the use of AdID, but were not explicitly asked by every app. By comparison, Google does not currently offer users the option to prevent apps from accessing the AdID on Android in general, but intends to change this from 'late 2021' [84]. The importance of consent aside, there exists little empirical evidence as to whether mobile apps implement any type of consent mechanisms before engaging in tracking.

Despite their crucial role within the software development life cycle, putting the blame of implementing consent incorrectly on app developers might be misguided. Many lack legal expertise, depend on the use of tracking software, and face limited negotiation power in the design of tracker software, which is usually developed by large, multinational companies [12, 21, 32, 49, 62]. At the same time, failure to implement

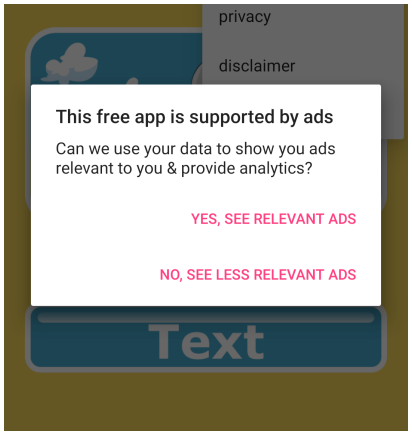
USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.



(a) This app uses the Consent API developed by Google. The popup suggests that personal data may be shared with 199 companies before user consent is given ('continue').



(b) This app uses the consent implementation by Twitter MoPub. By declining, a user rejects a 'personalized ad experience', but potentially not all app tracking.



(c) This app uses a custom consent solution. Consent is not granular. The answer options do not match the question. It is unclear if 'No' rejects analytics.

Figure 1: While most apps on the Google Play Store use third-party tracking, only few apps allow users to refuse consent (less than 3.5%). The figure shows three common examples of these 3.5% of apps. Since very few apps give users a genuine choice over tracking, our paper suggests widespread violations of EU and UK privacy law.

appropriate consent mechanisms in software impacts individuals' choice over data collection and their informational self-determination, and may expose vulnerable groups—such as children—to disproportionate data collection. This underlines the need for robust privacy guarantees in code.

Driven by these observations, the present contribution aims to answer the following research questions:

1. Do app developers need to obtain valid user consent before engaging in third-party tracking in the EU and UK? *(consent requirements for tracking)*
2. To what extent do apps engage in third-party tracking, and obtain valid user consent before doing so? *(practices of app developers)*
3. To what extent do third-party tracking companies encourage and support app developers to obtain consent as and where required? *(practices of tracker companies)*

Contributions. In answering these questions, this paper makes three contributions. First, we clarify the role of consent in the regulatory framework applicable in the EU and the UK when it comes to the processing of personal data for third-party tracking. Second, we provide empirical evidence as to a widespread absence of consent mechanisms to legitimise third-party tracking in 1,297 apps. Third, we analyse the guidance provided by 13 commonly used tracker companies and assess whether they inform app developers about how to translate consent in code (see Figure 2 and Table 2).

Structure. The rest of this paper is structured as follows. Section 2 reviews the relevant literature surrounding the concept of consent, app privacy analysis, and existing system-wide tracking controls for Android. Section 3 discusses the role of consent for third-party tracking in the EU and UK by drawing on the guidance issued by national Data Protection Authorities (DPAs). Section 4 analyses the presence of consent for third-party tracking in 1,297 Android apps randomly sampled from the Google Play Store. Section 5 reviews the guidance offered by tracker companies to app developers. After discussing the limitations of our approach in Section 6, we turn to the discussion of our results in Section 7 and our conclusions in Section 8.

2 Background

In this section, we discuss previous relevant literature, covering the concept of consent, the empirical analysis of privacy in apps, and existing system-wide tracking controls for Android. In particular, we highlight the limits of consent, and the dependence of end-users on the privacy options implemented by their apps and smartphone operating system.

2.1 Promises and Limits of Consent

Consent is a pillar of privacy and data protection law, in the US, EU, and many other jurisdictions and international frameworks. As an approach to privacy protection, consent is associated with the regime of *notice & choice* [74]. For many data-processing activities, companies that want to process data from an individual must

1. Adequately inform the individual (*Notice*), and

2. Obtain consent from the individual (*Choice*).

These two fundamental requirements are often implemented in software through the provision of a privacy policy, accompanied by consent options for the end-user.

The limitations of the notice & choice paradigm have been explored in a range of scholarship. Regarding "*notice*", it has been documented that most people do not read privacy policies, and that when they try to, have difficulties understanding them [69] and do not have enough time to read every such policy [61].

Regarding "*choice*", evidence suggests that many individuals struggle with privacy decisions in practice [5, 72]. The mismatch between stated and observed privacy preferences is known as the "*privacy paradox*" [64], although this so-called "*paradox*" may be best explained by structural forces that prevent alignment between values and behaviour [75, 82]. Individuals often have no real choice but to accept certain data processing because some digital services—such as Facebook or Google—have become indispensable [19]. Even when offered genuine choice, individuals face ubiquitous tracking [16], are tricked into consent [65], and do not get an adequate compensation in exchange for their data [23]. Because of the limits to individual privacy management, various scholars argue that the regime of notice & choice does not provide *meaningful* ways for individuals to manage their privacy [13, 14, 74].

Despite such limitations, consent remains a key component of many privacy and data protection regimes. For the purpose of this present contribution, we do not assume that consent is the only or best way to address privacy and data protection issues. Rather, we aim to investigate whether, in addition to all these problems and limitations, the basic process of consent itself is even being followed where it is currently required in the context of third-party tracking in apps.

2.2 Analysing Privacy in Apps

There is a vast range of previous literature that has analysed the privacy practices of mobile apps, and third-party tracking in particular. Two main methods have emerged in the academic literature: dynamic and static analysis.

Dynamic analysis executes an app, and analyses its run-time behaviour. While early research analysed apps by modifying the operating system [8, 33], recent work has focused on analysing apps' network traffic [50, 59, 66, 68, 70, 71, 73, 76, 80].

As for system modification, Enck et al. modified Android so that sensitive data flows through and off the smartphone could be monitored easily [33]. Agarwal and Hall modified iOS so that users were asked for consent to the usage of sensitive information by apps [8], before the introduction of run-time permissions by Apple in iOS 6.

As for network analysis, Ren et al. instrumented the VPN functionality of Android, iOS, and Windows Phone to expose leaks of personal data over the Internet [70]. Conducting a manual traffic analysis of 100 Google Play and 100 iOS apps, they found regular sharing of personal data in plain text, including device identifiers (47 iOS, 52 Google Play apps), user location (26 iOS, 14 Google Play apps), and user credentials (8 iOS, 7 Google Play apps). Van Kleek et al. used dynamic analysis to expose unexpected data flows to users and design better privacy indicators for smartphones [80]. Reyes et al. used dynamic analysis to assess the compliance of children's apps with COPPA [71], a US privacy law to protect children. Having found that 73% of studied children's apps transmit personal data over the Internet, they argued that none of these apps had obtained the required "*verifiable parental consent*" because their automated testing tool could trigger these network calls, and a child could likely do so as well. Okoyomon et al. found widespread data transmissions in apps that were not disclosed in apps' privacy policies, and raised doubts about the efficacy of the notice & choice regime [66] (as discussed in the previous section).

Dynamic analysis offers different advantages. It is relatively simple to do, largely device-independent, and can be used to monitor what data sharing actually takes place. It has, however, several limitations. The information gathered might be incomplete if not all code paths within the app involving potential data disclosures are run when the app is being analysed. Further, network-based dynamic analysis may wrongly attribute system-level communications to a studied app, e.g. an Android device synchronising the Google Calendar in the background, or conducting a network connectivity check with Google servers. Network-based dynamic analysis remains nonetheless a versatile, reliable and practical approach.

Static analysis infers the behaviour of an app without the need for execution. This process often relies on decompiling an app and analysing the retrieved program code [31, 51]. The main advantage of static analysis is that it enables the analysis of apps at a much larger scale (e.g. millions rather than hundreds) [16, 20, 81, 83]. As opposed to dynamic analysis, static analysis may require substantial computing resources and does not permit the direct observation of network traffic because apps are never run.

Egele et al. developed an iOS decompiler and analysed 1,407 iOS apps. They found that 55% of those apps included third-party tracking libraries [31]. Viennot et al. analysed more than 1 million apps from the Google Play Store, and monitored the changing characteristics of apps over time [81]. They found a widespread presence of third-party tracking libraries in apps (including Google Ads in 35.73% of apps, the Facebook SDK in 12.29%, and Google Analytics in 10.28%). Similarly, Binns et al. found in analysing nearly 1 million Google Play apps that about 90% may share data with Google, and 40% with Facebook [16].

The presence of consent to tracking in apps has received

relatively little research attention; to the best of our knowledge, no large-scale studies in this area exist. With static analysis, it is difficult to detect at what stage a user might give consent, because of the varied implementations of consent in app code. However, network-based dynamic analysis makes this kind of consent analysis possible, and at a reasonable scale. We demonstrate this in Section 4.

2.3 Alternatives to In-App Consent

Before turning to the legal analysis concerning when consent for third-party tracking within individual apps is required, it is worth considering the options users currently have to limit app tracking on Android at a system level. This is pertinent to our subsequent analysis because, if system-level controls were sufficient, the question of efficacy and compliance with individual app-level consent requirements might be redundant. The options for users fall into three categories: system settings, system modification, and system APIs.

System Settings. The Android operating system offers users certain possibilities to limit unwanted data collection. Users can manage the types of data each app can access through *permissions*. This does not stop tracking, but blocks access to certain types of data, such as location. A problem inherent to the permission approach is that trackers share permission access with the apps they come bundled with. This means that, if a user allows location access to a maps app with integrated trackers, all these trackers have access as well. This, in turn, might give users a false sense of security and control. Google offers users the possibility to opt-out from personalised advertising. If users choose to do so, apps are encouraged to cease using the system-wide *Google Advertising Identifier* (AdID) for personalised advertising (although apps can continue to access the AdID). Unlike iOS, Android does yet not offer the option to opt-out from analytics tracking using the AdID, or to prevent apps from accessing this unique user identifier. However, Google intends to change this from 'late 2021' [84].

System Modification. Since the early days of Android, many developers have set out to modify its functionality and implement better privacy protections. *Custom ROMs* are modified versions of Android that replace the default operating system that comes pre-installed on Android smartphones. Popular examples are Lineage OS and GrapheneOS, which both try to reduce the dependency on Google on Android and increase user privacy. Another is TaintDroid, which monitors the flow of sensitive information through the system [33]. A popular alternative to custom ROMs is *rooting* devices by using exploits in the Android system to gain elevated access to the operating system, or by changing the bootloader of the Android system. Rooting is a necessary prerequisite for many privacy-focused apps, including AdAway [6], XPrivacy [18], and AppWarden [11]. System modification grants maximum control and flexibility regarding tracking, but requires a high level of technical expertise. It also relies on security vulnerabilities, often creating risks for (non-expert) users.

As a result, Google has recently begun to restrict attempts to modify Android by preventing custom ROMs from running apps using *Google's Safety Net*. This is meant to protect sensitive apps (e.g. banking apps) from running on unsafe devices, but is also used by other popular apps such as Pokemon GO and Snapchat [41]. Some Internet outlets have declared the "*end for Android rooting, [and] custom ROMs*" [77].

System APIs. Another alternative to system modification is to develop apps that build on the capabilities of Android's system APIs to detect and block network traffic related to tracking. Such is possible without the need for system modification at the cost of more advanced functionality. Popular apps in this category include AdGuard (using a local VPN on the Android device) [7] and DNS66 (changing the DNS settings of the Android device) [57]. Another is NetGuard [17], a firewall that allows users to monitor network connections through a VPN, and to block certain domains manually. All these tools block connections regardless of the actual content of the communications. These content-agnostic approaches can lead to overblocking and break apps.

Alternative tools aim for more fine-grained protection by removing sensitive information from network requests, such as device identifiers or location data [59, 76]. Unfortunately, these content-based approaches rely on breaking secured network connections, and on installing a self-signed root certificate on the user's device. This practice was banned by Google with the introduction of Android 7 in 2016 because of the security risks it entails [44]. While these apps grant users the possibility to block tracking through system APIs, Google does not allow them on the Play Store [45]. Instead, users must sideload them onto their device from alternative sources, such as GitHub and F-Droid.

In conclusion, while there exists a wide array of options for end-users to reduce tracking, none of them can provide the granularity of consent implemented inside each individual app. Many of the existing tools require a high level of technical expertise, including root access or modifications to the operating system, and are therefore unsuitable for non-expert users. This makes many users dependent on the privacy solutions offered by apps themselves and their operating systems.

3 When is Consent to Tracking required?

In this section, we analyse whether consent is a prerequisite for third-party tracking under EU and UK law, as well as its role under the Google Play Store policy. We focus on these jurisdictions as they have relatively stringent and specific rules on consent and third-party tracking. While similar rules exist in other jurisdictions (such as the COPPA in the US, which requires parental consent for tracking), recent regulatory actions and rich guidance issued by European regulators offer an ideal setting for a large-scale analysis.

Two main legal instruments are relevant to the issue of consent to third-party tracking on mobile apps: the GDPR and the ePrivacy Directive[1].

3.1 GDPR and the Need for a Lawful Ground

Applicable since 25 May 2018, the GDPR grants users various rights over their *personal data*. It also imposes a wide array of obligations on so-called *controllers*, paired with high fines for non-compliance. One of the cornerstones of the legislative reform is the concept of *Data Protection by Design*; this obliges controllers to implement appropriate technical and organisational measures to ensure and demonstrate compliance with all rules stemming from the Regulation throughout the entire personal data processing life cycle (Article 24(1) and 25(1) GDPR). All companies operating in the EU or UK must follow the rules set out in the GDPR.[2] Compliance with the GDPR is monitored and enforced by the *Data Protection Authorities* (DPAs) instituted in each EU Member State and in the UK. If a controller fails to comply, DPAs have the power to impose fines that can go up to €20 million or 4% of the company's worldwide annual turnover, whichever is higher.

For the purpose of this paper, we assume that app developers qualify as *controllers*. In other words, that they "*determine the purposes and the means of the processing of personal data*" (Article 4(7) GDPR). While this might well be the case when the company actually processing the personal data at stake is also in charge of the development of the app, it is important to highlight that controllership does not always end up on their shoulders. This is the case, for instance, when a company outsources the development of its app to an external team of software developers working on the basis of clear-cut specifications and requirements, in which case the latter is likely to be considered as a *processor* (Article 4(8) GDPR) or a *third party* (Article 4(10) GDPR).

If app developers want to collect personal data for whatever purpose, they need to rely on one of the six *lawful grounds* listed in Article 6(1) GDPR. Only two usually apply in the context of mobile apps, namely: *consent* and *legitimate interests*[3]. On the one hand, and as specified in Article 4(11) GDPR, a valid *consent* is

> *any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her*

As recently clarified by the Court of Justice of the European Union, this bans the use of pre-ticked boxes to gather consent [27]. *Legitimate interests*, on the other hand, is a viable alternative to consent but requires a careful balancing exercise between the controller's interests in processing the personal data and the data subjects' interests and fundamental rights (Article 6(1)f GDPR) [35]. The other guarantees stemming from the GDPR (including transparency, security, purpose and storage limitation, and data minimisation) remain applicable regardless of the lawful ground used to legitimise the processing.

Consent for High-Risk Data Processing. While the controller's legitimate interests could potentially be a viable option for legitimising third-party tracking on mobile apps, this processing is likely to qualify as a *high-risk data processing activity*.[4] Features of third-party tracking that indicate such high-risk processing include the use of "*evaluation or scoring*", "*systematic monitoring*", "*data processed on a large scale*", "*data concerning vulnerable data subjects*", or "*innovative use or applying new technological or organisational solutions*". Some of these features undoubtedly apply to third-party tracking, since tracking companies usually engage in large-scale data collection, at a high-frequency, across different services and devices, with limited user awareness.

The Information Commissioner's Office (ICO)—the UK's DPA—discourages the use of legitimate interest for high-risk activities and recommends that controllers instead rely on another lawful ground such as consent [52]. Similarly, after having analysed the case of tracking deployed on a webshop selling medical and cosmetic products [28], the German DPA came to the conclusion that the average website visitor could not reasonably expect tracking of their online activities to take place, especially when it operates across devices and services. In that case, it argued, the website visitor is not in a position to avoid the data collection. These are the concrete manifestations of the balancing exercise required by Article 6(1)f. All in all, the above-mentioned considerations disqualify the use of the controllers' legitimate interests as an appropriate lawful ground to legitimise third-party tracking in mobile apps.

[1] It is worth noting that the ePrivacy Directive is currently under revision. A change to the current regulatory requirements in practice is expected no earlier than in two years due to the nature of the EU legislative process.

[2] More specifically, all companies based in the EU and UK, as well as companies monitoring the behaviour of, or offering goods and services to, individuals located in the EU and UK, fall within the territorial scope of application of the GDPR (Article 3 GDPR).

[3] The remaining four lawful grounds listed in Article 6(1) GDPR being the fulfilment of a *contract*, a *legal obligation*, the data subject's *vital interests*, and the performance of a *public task*.

[4] The Article 29 Working Party—an EU body to provide guidance on data protection law (now the European Data Protection Board)—has listed the 9 features commonly found in such high-risk activities, namely: 1) Evaluation or scoring, 2) Automated-decision making with legal or similar significant effect, 3) Systematic monitoring, 4) Sensitive data or data of a highly personal nature, 5) Data processed on a large scale, 6) Matching or combining datasets, 7) Data concerning vulnerable data subjects, 8) Innovative use or applying new technological or organisational solutions, and 9) Prevention of data subjects from exercising a right or using a service or a contract. [36]

3.2 ePrivacy and the Need for Consent for Local Storage of and Access to Data

In addition to the GDPR, the ePrivacy Directive also applies to third-party tracking. This is a *lex specialis*, meaning that, when both the ePrivacy Directive and the GDPR apply in a given situation, the rules of the former will override the latter. This is the case for third-party tracking, since Article 5(3) of the ePrivacy Directive specifically requires consent for accessing or storing non-technically necessary data on a user's device. It is widely accepted, and reflected in DPAs' guidance, that most tracking activities are not technically necessary, and therefore require consent to store data on a user's device [54]. So, if tracker software involves accessing or saving information on a user's smartphone—as third-party trackers typically do on a regular basis—this requires prior consent. As a result, while consent was already the most reasonable option under the GDPR, it becomes the only viable one when combining both regulatory frameworks.

As stated above, the GDPR provides a range of possible lawful grounds of which consent is just one; however, Article 5(3) of the ePrivacy Directive specifically requires consent for accessing or storing non-technically necessary data on a user's device. As a consequence, any further processing by the third party which is not covered by the initial consent interaction would usually require the third party to obtain fresh consent from the data subject.

Recent guidance and enforcement action from various DPAs have also demonstrated how the GDPR and the ePrivacy requirements apply to situations where consent is the basis for processing by one controller, and when that data is provided to another controller for further processing. Article 7(1) of the GDPR requires that, where consent is the lawful ground, the controller must be able to *demonstrate* that the data subject has consented. The ICO's guidance states that third-party services should not only include contractual obligations with first parties to ensure valid consent is obtained, but "*may need to take further steps, such as ensuring that the consents were validly obtained*" [53]. It notes that, while the process of getting consent for third-party services "*is more complex*", "*everyone has a part to play*" [53]. The responsibility of third parties has been further illustrated in an enforcement action by the CNIL (the French DPA), against Vectaury, a third-party tracking company [22]. This showed how the validity of consent obtained by an app developer is not "*transitive*", i.e. does not carry over to the third party. If a first party obtains consent "*on behalf*" of a third party, according to a contract between the two, the third party is *still* under the obligation to verify that the consent is valid.

To summarise the implications of GDPR and ePrivacy in the context of third-party tracking: consent is typically required for access to and storage of data on the end-user's device. Even if that consent is facilitated by the first party, third parties must also be able to demonstrate the validity of the consent for their processing to be lawful on that basis.

3.3 Requirements of the Google Play Store

In addition to EU and UK privacy law, Google imposes a layer of contractual obligations that apps must comply with. These policies apply worldwide—so beyond the jurisdiction of the EU and UK—and might oblige all app developers to implement adequate mechanisms to gather consent for third-party tracking. Google's *Developer Content Policy* highlights that in-app disclosure and consent might need to be implemented when "*data collection occurs in the background of your app*" [47]. The Developer Content Policy also requires that developers abide by all applicable laws. It is unclear how strictly compliance with these policies—and in particular with all applicable laws—is verified and enforced by Google.

4 Tracking in Apps Before and After Consent

The previous section established that third-party tracking in apps requires valid user consent under the EU and UK regulatory framework. Despite these legal obligations, it yet not clear how and whether consent is realised in practice. In order to examine the extent to which regulation around consent is implemented in practice, we conducted two studies—Study 1 (in this section) to see how consent is implemented in a representative sample of Google Play apps, and Study 2 (in the following Section 5) to examine how app developers were supported and encouraged to implement consent by the providers of tracker libraries.

4.1 Methodology

We studied a representative sample of 1,297 free Android apps from the UK Google Play Store. This sample was chosen randomly (through random sampling without replacement) from a large set of 1.63 million apps found on the Google Play Store between December 2019 and May 2020 to understand the practices across the breadth of app developers. We explored the presence of apps on the app store by interfacing with Google Play's search function, similar to previous research [81]. The selected apps were run on a Google Pixel 4 with Android 10. Each app was installed, run for 15 seconds, and then uninstalled. We did not interact with the app during this time, to record what companies the app contacts before the user can be informed about data collection, let alone give consent. During app execution, we recorded the network traffic of all tested apps with the popular NetGuard traffic analysis tool [17]. We did not include any background network traffic by other apps, such as the Google Play Services. For apps that showed full-screen popup ads, we closed such popups, and took note of the presence of display advertising. We assessed whether each contacted domain could be used for tracking

Hosts	Company	Apps
adservice.google.com	Alphabet	19.7%
tpc.googlesyndication.com	Alphabet	17.2%
lh3.googleusercontent.com	Alphabet	14.2%
android.googleapis.com	Alphabet	12.9%
csi.gstatic.com	Alphabet	11.6%
googleads.g.doubleclick.net	Alphabet	10.3%
ade.googlesyndication.com	Alphabet	9.7%
connectivitycheck.gstatic.com	Alphabet	9.5%
config.uca.cloud.unity3d.com	Unity	7.5%
ajax.googleapis.com	Alphabet	6.9%
api.uca.cloud.unity3d.com	Unity	6.8%
android.clients.google.com	Alphabet	6.7%
gstatic.com	Alphabet	5.8%
graph.facebook.com	Facebook	5.5%

Table 1: Top contacted tracker domains by 1,201 randomly sampled apps from the Google Play Store, at launch, before any interaction with the apps.

and, if so, to what tracking company it belonged, using a combination of the App X-Ray [15] and `Disconnect.me` [29] tracker databases. 15 seconds after having installed the app, we took a screenshot for further analysis, and uninstalled it.

We inspected the screenshots for any form of display advertising, privacy notice or consent. We took note of any display advertising (such as banner and popups advertising) observed. We classified any form of information about data practices as a privacy notice, and any *affirmative* user agreement to data practices as consent. While this definition of consent is arguably less strict than what is required under EU and UK law, this was a deliberate choice to increase the objectivity of our classification, and provide an upper bound on compliance with EU and UK consent requirements. We then re-installed and ran those apps that asked for consent, granted consent, and repeated the network capture and analysis steps above, i.e. monitoring network connections for 15 seconds, followed by a screenshot, and finally, removed the app once again.

4.2 Results

Of the 1,297 apps, 96 did not show a working user interface. Some apps did not start or showed to be discontinued. Other apps did not provide a user interface at all, such as widgets and Android themes. We therefore only considered the remaining 1,201 apps. 909 apps (76%) were last updated after the GDPR became applicable on 25 May 2018.[5] On average, the considered apps were released in August 2018 and last updated in December 2018. All apps were tested in August 2020, within a single 24-hour time frame.

Widespread tracker use. Apps contacted an average of 4.7 hosts each at launch, prior to any user interaction. A majority of such apps (856, 71.3%) contacted known tracker hosts. On average, apps contacted 2.9 tracker hosts each, with a standard deviation of 3.5. The top 10% of apps contacted at least 7 distinct hosts each, while the bottom 10% contacted none. Alphabet, the parent company of Google, was the most commonly contacted company (from 58.6% of apps), followed by Facebook (8.2%), Unity (8.2%), One Signal (5.6%), and Verizon (2.9%). Apps that we observed showing display ads contacted a significantly higher number of tracker hosts (on average 6.0 with ads vs 2.2 without).

Dominance of Google services. The 9 most commonly contacted domains all belong to Google; the top 2 domains are part of Google's advertising business (`adservice.google.com`, linked to Google's Consent API, and `tpc.googlesyndication.com`, belonging to Google's real-time advertisement bidding service). 704 apps (58.6%) contacted at least one Google domain; the top (Google) domain was contacted by 236 apps (19.7%). Such breadth and variation is reflective of the corresponding variety of services that Google offers for Android developers, including an ad network (Google AdMob), an ad exchange (Google Ad Manager, formerly known as DoubleClick), and various other services. Domains by other tracker companies, such as Unity and Facebook, were contacted less frequently by apps (see Table 1).

Google's tracking was also observed to be deeply integrated into the Android operating system. It has been known that the Google Play Services app—required to access basic Google services, including the Google Play Store—is involved in Google's analytics services [63]. In our network analysis, this app seemed to bundle analytics traffic of other apps and send this information to Google in the background with a time delay. Without access to encrypted network traffic (as explained in Section 2.3), this makes it impossible to attribute network traffic to individual apps from our sample, when such network traffic could also be related to other system apps (some of which, such as the Google Phone app, use Google Analytics tracking themselves). As a consequence, we are likely under-reporting the number of apps that share data with Google, since we only report network traffic that could be clearly attributed.

Consent to tracking is widely absent. Only 9.9% of apps asked the user for consent. Apps that did so contacted a larger number of tracker hosts than those that did not (3.7 with consent vs 2.8 that did not). A slightly larger fraction (12.2% of all apps), informed the user to some extent about their privacy practices; apps in this category also contacted a larger number of trackers than those that did not (3.6 that informed vs 2.8 that did not). 19.1% of apps that did not ask for consent showed ads, compared to only 2.5% of apps that asked for consent. Once consent was granted, the apps

[5] It is worth noting, however, that both the need for a lawful ground—an obligation under Directive 95/46—and the consent requirement for access to and storing on terminal equipment—an obligation under the ePrivacy Directive—were already applicable before 25 May 2018. The latter has merely provided clarification on the conditions for consent to be valid.

contacted an average of 4.2 tracker hosts (higher than the 3.7 before granting consent, and the 2.8 for apps without any consent flows).

Consent limited to using or not using an app. Most apps that ask for consent force users into granting it. For instance, 43.7% of apps asking for consent only provided a single choice, e.g. a button entitled "*Accept Policy and Use App*" or obligatory check boxes with no alternative. A further 20.2% of apps allowed users to give or refuse consent, but exited immediately on refusal, thus providing a *Hobson's choice*. Only 42 of the apps that implemented consent (comprising a mere 3.5% of all apps) gave users a genuine choice to refuse consent. However, those apps had some of the highest numbers of tracker hosts, and contacted an average of 5.2 on launch. Among these apps, if consent was granted, the number of tracker hosts contacted increased to 8.1, but, interestingly, an increase was also observed even if data tracking was opted-out (from the pre-consent 5.2 to 7.5 post-opt-out).

Consent limited to the personalisation of ads. Consent was often limited to an opt-out from personalised ads. 37 of the 42 apps that implement a genuine choice to refuse consent restrict this choice to limiting personalised advertising; such choice might make some users wrongly assume that refusing to see personalised ads prevents all tracking (see Figure 1 for some common examples). We observed that 23 of these 37 apps (62%; 1.9% overall) used Google's Consent API [43], a toolkit provided by Google for retrieving consent to personalised ads (particularly when multiple ad networks are used). None of the apps using the Google Consent API, however, ended up asking users to agree to further tracking activities, such as analytics. Only 4 apps provided the option to refuse analytics; all 4 of these did so in addition to providing the option to opt-out of personalised advertising. One further app in our sample requested consent to process health data.

5 Support and Guidance from Trackers

The previous section found a widespread absence of consent to third-party tracking in apps. As explained in Section 3, both first and third parties have a part to play in facilitating valid consent, and third parties need to take steps to ensure consent obtained by first parties is valid. At the same time, it has been reported that many app developers believe the responsibility of tackling risks related to ad tracking lie with the third-party companies [62], and need clear guidance regarding app privacy [12]. In this section, we assess the efforts, that providers of tracker libraries make, to encourage and support app developers in implementing a valid consent mechanism. We focus on the most common libraries so as to understand the current practices across the tracking industry.

5.1 Methodology

Our qualitative analysis focuses on the 13 most common tracker companies on Android (according to [39]), and three types of document that each of them provides: 1) a step-by-step implementation guide, 2) a privacy policy, and 3) further publicly available documentation. While there may be other ways in which providers of tracking libraries support app developers to facilitate valid consent, we reason that these are the standard means by which such support would be provided. Step-by-step implementation guides serve as a primary resource for app developers and summarise the essential steps of implementing a tracker library in code. Since the implementation of consent must be done in code, consent implementation is one essential step for those trackers that require consent.

In assessing this documentation, we assume the perspective of an app developer who is motivated to comply with any explicit requirements mentioned by the tracker provider, and to follow their instructions as to how to do so, but lacks in-depth knowledge about how the GDPR and ePrivacy Directive apply to their use of a given third-party tracking software [49]. We also assume that app developers are likely to read documentation only so far as necessary to make the third-party library functional, often through trial-and-error [56, 58], and stop studying other resources once the tracker implementation is functional, since they are often pressured by time and economic constraints [4, 32, 62].

5.2 Results

Our results are summarised in Table 2. We detail our main findings in the following paragraphs.

Most trackers are unclear about their use of local storage. Whether a tracker accesses and/or stores information on a user's device is essential in determining the need to implement consent, as explained in Section 3.2. As such, we would expect to find information stating whether or not access and/or storage takes place as part of the standard operation of the tracker. However, we did not find such information for 6 out of 13 trackers. For the others, this information was difficult to find. AppsFlyer rightly states in its online documentation that "*there are no cookies for mobile apps or devices*" [9]. While this is true from a technical perspective, EU and UK law do not differentiate between cookies and other information saved on a user's device. Crucially, we did not find any tracker stating *not* to save information on a user's device. In the absence of such a denial, app developers would run the risk of assuming they do not need to obtain consent for data accessed and/or stored by the tracker.

Most trackers expect app developers to obtain consent. Despite being unclear about their use of local storage, a closer inspection of the tracker policies and documentation found that most trackers instruct developers to request consent from EU users (11 out of 13). AppLovin is an exception, but does

Tracker	Apps	Expects consent (in EU / UK)	Implements consent (by default)	Mentions consent (in implementation guide)	Discloses local data storage
Google Analytics	50%	Yes	No	No	Yes
Google AdMob	45%	Yes	No	Yes	Yes
Google Crashlytics	29%	Yes	No	No	Yes
Facebook App Events	20%	Yes	No	No	?
Google Tag Manager	19%	Yes	No	No	Yes
Facebook Ads	14%	Yes	Yes*	No	?
Flurry	9%	Yes	No	No	?
Unity Ads	8%	Yes	Yes	No	Yes
Inmobi	8%	Yes	No	Yes	?
Twitter MoPub	6%	Yes	Yes	No	Yes
AppLovin	6%	No	No	No	?
AppsFlyer	5%	?	No	Yes	?
OneSignal	4%	Yes	No	No	Yes

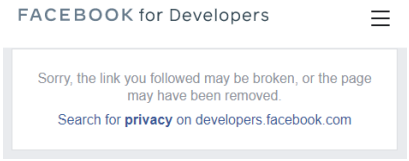
Table 2: Consent requirements and implementation for 13 commonly used Android trackers. App shares according to the Exodus Privacy Project [39]. The **trackers in bold** require consent, but do neither implement such by default nor mention the need to do so in their implementation guides. ?: We did not find any information. *: Facebook opts-in users by default to their personalised advertising, unless they disable this behaviour from their Facebook settings or do not use the Facebook app.

require consent if developers want to show personalised ads (which tend to be more lucrative than contextual ads). For AppsFlyer, we could not find any information regarding the need to ask users for consent. The need to ask for consent was sometimes difficult to find, and required a careful reading of the policies and documentation provided. Some developers are bound to overlook this, and unnecessarily compromise on the users' right to choose over tracking.

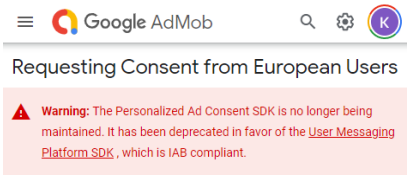
Few trackers implement consent by default. We further inspected whether tracker libraries provide their own consent implementation. If they do, an app developer would not need to make any further modification to the app code. However, only a minority of tracker libraries (3 out of 13) integrates an implementation of user consent by default, and none of the five most common trackers do so. Unity Ads and Twitter MoPub provide consent flows that are automatically shown, without further action by the app developer. Facebook Ads only shows ads, if the app user 1) has agreed to personalised ads in their Facebook account settings, and 2) uses the Facebook app on their phone. However, Facebooks opts-in users by default to their personalised advertising, unless they disable this behaviour from their Facebook settings (checked 14 February 2021). While Google AdMob provides a consent library, this is not implemented by default. Indeed, Google AdMob expects the app developer to retrieve consent from the user, but shows personalised ads even if the developer does not implement their consent library.

Limited disclosure of consent requirements in step-by-step guides. We find that 3 out of 13 tracker libraries disclose the potential need for consent in their step-by-step implementation guides. This is despite 11 out of 13 trackers mentioning the need to implement consent in other places of their online documentation. Google AdMob mentions the need to retrieve consent amongst other "*examples of actions that might be needed prior to initialization*" [46] of AdMob. Inmobi points out that developers need to "*obtain appropriate consent from the user before making ad requests to InMobi for Europe*" [55] in the Section on "*Initializing the SDK*". AppsFlyer offers developers to "*postpone start [of the tracker library] until you receive user consent due to GDPR or CCPA requirements, etc.*" [10] in Section 3.4 on "*Delay SDK initialization*". It is not clear from these three implementation guides what other reasons are to "*delay initialisation*" beyond legal compliance, and why this is not clarified. At least 6 out of 13 trackers require consent, but neither implement such by default nor inform app developers of the need to do so in the implementation guides. If AppLovin needs consent (despite not stating to do so, but as suggested by our legal analysis in Section 3), this figure would increase to 7 out of 13 trackers.

Compliance guidance: often provided, but sometimes difficult to find, hard to read, and poorly maintained. Many tracker companies provide additional information on GDPR compliance and consent implementation on a separate website as part of their online documentation. We found some compliance guidance (with varying levels of detail) for all trackers except the Google Tag Manager. Excluding the 3 trackers implementing consent by default, a developer needs an average of 1.56 clicks to reach these compliance guides. For AppLovin, a developer must click "*Help Center*", then "*Getting started & FAQ*", and lastly "*User opt-in/opt-out in the AppsFlyer SDK*". Facebook required developers to click "*Best Practices Guide*" and then "*GDPR Compliance guide*".



(a) Facebook links to a page that is supposed to explain how to implement consent in practice.



(b) Google AdMob links to an outdated library, creating unnecessary friction for consent implementation.

(c) Flurry links to a broken GDPR guide.

Figure 2: Many trackers provide information on what developers need to know to implement consent. These guides are often difficult to find, hard to read, and poorly maintained. 3 out of 13 common trackers linked to unmaintained or broken pages.

While this GDPR compliance guide provides some guidance on the implementation of consent, the link to Facebook's "*consent guide*" with practical examples of how to implement consent was broken. Also, the framing as "*Best Practices*" suggests optionality of legal compliance. For OneSignal, developers must first click "*Data and Security Questions*" and then "*Handling Personal Data*".

The compliance guides (excluding code fragments) reached a mean Flesch readability score [42] of 41.8, as compared to 50.6 for the step-by-step implementation guides (where 100 means "*very easy*", and 0 "*very difficult*" to read). Both the implementation and compliance guides are "*difficult*" to read, with the compliance guides somewhat more so. For 3 of the 13 trackers, we were directed to broken or outdated links (see Figure 2). Google AdMob linked to an outdated consent strategy, while the Facebook SDK and Flurry linked to non-existing pages (returning `404` errors). We found other pages with compliance information for each of these trackers, but broken guidance can act as a deterrent for developers who want to implement consent and follow their legal obligations. However, while this paper was under review, the broken links in the documentation of the Flurry and Facebook trackers were fixed.

6 Limitations

It is important to acknowledge some limitations of our methodology. Our analysis in Section 4 used dynamic analysis, and not all tracking might be detected. We only inspected network traffic before and shortly after consent was given. Apps might therefore conduct more tracking during prolonged app use. Besides, we only reported the network traffic that could be clearly attributed to one of the apps we studied, potentially leading to under-reporting of the extent of Google's tracking (as explained in Section 4). While the reported tracking domains can be used for tracking, they might also be used for other non-tracking purposes; however, it is the choice of the tracking company to designate domains for tracking. We do not study the contents of network traffic because apps increasingly use certificate pinning (about 50% of the studied apps used certificate pinning for some of their network communications). As for our second study in Section 5, we studied the online documentation of tracker libraries with great care, but did not always find all relevant information, particularly regarding the local storage of data on a user's device. Where this was the case, we disclosed this (e.g. see Table 2).

7 Discussion and Future Work

Consent is an integral part of data protection and privacy legislation, both in the EU and the UK, and elsewhere. This is all the more so in the context of third-party tracking, for which consent appears the only viable lawful ground under the ePrivacy Directive and the GDPR, as analysed in Section 3. Not only has this has been emphasised by multiple DPAs, but is also acknowledged by tracking companies themselves in the documentation they make available to app developers. Relying on the controller's legitimate interests—the only conceivable alternative to consent under EU and UK data protection law—would likely fail short of passing the balancing test outlined in Article 6(1)f GDPR. This also follows from the requirement to obtain consent prior to storing or accessing information on a user's device, under the ePrivacy Directive.

Against this backdrop, we analysed 1,297 mobile apps from Google Play in Section 4 and discovered a widespread lack of appropriate mechanisms to gather consent as required under the applicable regulatory framework. We found that, while the guidelines of many commonly used tracker libraries require consent from EU and UK users, most apps on the Google Play Store that include third-party tracking features do not implement any type of consent mechanism. The few apps that require data subjects to consent do so with regard to personalised advertising, but rarely for analytics—despite this being one of the most common tracking practices. Where an opt-out from personalised advertising was possible, the number of tracker domains contacted decreased only slightly after opting-out, hinting to continued data collection when serving contextual advertising. These observations are at odds with the role of consent as the only viable option to justify the

processing of personal data inherent to third-party tracking.

As detailed in Section 4, the fact that only 9.9% of the investigated apps request any form of consent already suggests widespread violations of current EU and UK privacy law. This is even before considering the validity of the consent mechanisms put in place by that small fraction of apps. As underlined in Section 3, consent must be "*freely given*", "*informed*", "*specific*" and "*unambiguous*". The findings outlined in Section 4 suggest that most apps that do implement consent force users to grant consent, therefore ruling out its qualification as "*freely given*". The same goes for the 43.7% of those apps that do not provide data subjects with the possibility to consent separately for each purpose, but instead rely on *bulk* consent for a wide array of purposes.

When considering both the absence of any form of consent in more than 90% of the investigated apps and the shortcomings inherent to the few consent mechanisms that are implemented by the remaining sample, we infer that the vast majority of mobile apps fail short of meeting the requirements stemming from EU and UK data protection law. Our analysis does not even consider the fact that consent is only one of a variety of legal rules that third-party tracking needs to comply with. Breaches of other legal principles—such as data minimisation, purpose and storage limitation, security and transparency—might be less visible than a lack of consent and harder to analyse, but no less consequential.

We further found that one of the reasons for the lack of consent implementation in apps might be inadequate support by tracker companies [12, 62]. Studying the online documentation of the 13 most commonly used tracker libraries in Section 5, only 3 trackers implemented consent by default, and another 3 disclosed the need to implement consent as part of step-by-step implementation guides. These step-by-step guides serve as a primary resource for app developers, and can give a false impression of completeness when in fact additional code needs to be added for many trackers to retrieve user consent. This is true for at least 6 out 13 trackers, including Google Analytics and the Facebook App Events SDK, which likely need consent, but neither disclose this in their implementation guides nor implement such consent by default. While most trackers provide some compliance guidance, we found that this can be difficult to find, hard to read, and poorly maintained. Whatever the reasons for the lack of consent, the result is an absence of end-user controls for third-party tracking in practice.

Lastly, it is worth highlighting that Google, which is both the largest tracking company and the main developer of Android, faces conflicts of interest with respect to protecting user privacy in its Google Play ecosystem [30, 48, 78]. The company generates most of its revenue from personalised advertising, and relies on tracking individuals at scale. Certain design choices by Google, including its ban of anti-tracking apps from the Play Store, its recent action against modified versions of Android, and the absence of user choice over AdID access for analytics on Android (as opposed to iOS), create friction for individuals who want to reduce data collection for tracking purposes, and lead to increased collection of personal data, some of which is unlawful as our legal analysis has shown.

Future work. An overarching question for future work is the extent of the legal obligations faced by the many actors involved in the third-party tracking ecosystem, ranging from app developers to providers of tracker libraries and mobile operating systems. This is inextricably linked to their qualification as "*controllers*", a legal notion the boundaries of which remain, despite recent jurisprudence [24–26] and detailed guidance [34, 37], still controversial. Our analysis highlighted how simple changes in the software design can have significant effects for user privacy.

Moreover, while the US—unlike many developed countries—lack a federal privacy law, there exists a variety of specific privacy laws, such as COPPA to protect children and HIPAA to protect health data, as well as state-level privacy laws, including CCPA in California. Some of these laws foresee consent requirements similar to EU and UK law. We leave it to further work to assess how widely apps comply with the consent requirements of US privacy legislation.

8 Conclusions

Our work analyses the legal requirements for consent to tracking in apps, and finds an absence of such consent in practice based on an analysis of a representative sample of Google Play apps. This, in turn, suggests widespread violations of EU and UK privacy law. Simple changes by software intermediaries (such as Google and Facebook), including default consent implementations in tracker libraries, better legal guidance for app developers, and better privacy options for end-users, could improve the status quo around app privacy significantly. However, there is doubt that these changes will happen without further intervention by independent parties—not only end-users, but also policymakers and regulators—due to inherent conflicts between user privacy and surveillance capitalism.

While the web has seen a proliferation of deceptive and arguably meaningless consent banners in recent years [60, 65], we hope that mobile apps will not see a similar mass adoption. Rather, we aim to influence the current policy discourse around user choice over tracking and ultimately to make such choice more meaningful. As Apple has demonstrated with its recently introduced iOS 14.5, system-level user choices can standardise the process of retrieving user consent and make hidden data collection, such as tracking, more transparent to end-users. We call on policy makers and data protection regulators to provide more stringent guidelines as to how consent to tracking should be implemented in program code, particularly by browser developers, device manufacturers, and platform gatekeepers in the absence of such existing requirements.

References

[1] 9to5mac.com. Apple rebuffs Facebook criticism, says iOS anti-tracking features are about 'standing up for our users'. https://9to5mac.com/2020/12/16/apple-facebook-app-tracking-transparency/, 2020.

[2] 9to5mac.com. Facebook attacks Apple in full-page newspaper ads. https://9to5mac.com/2020/12/16/facebook-attacks-apple/, 2020.

[3] 9to5mac.com. Apple versus Facebook on ad-tracking: Harvard sides with Apple. https://9to5mac.com/2021/02/05/apple-versus-facebook-harvard/, 2021.

[4] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L. Mazurek, and Christian Stransky. You get where you're looking for: The impact of information sources on code security. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 289–305, 2016.

[5] Alessandro Acquisti. Nudging Privacy: The Behavioral Economics of Personal Information. *IEEE Security & Privacy Magazine*, 7(6):82–85, 2009.

[6] AdAway. AdAway. https://github.com/AdAway/AdAway, 2021.

[7] AdGuard. AdGuard for Android. https://adguard.com/en/adguard-android/overview.html, 2021.

[8] Yuvraj Agarwal and Malcolm Hall. ProtectMyPrivacy: Detecting and mitigating privacy leaks on iOS devices using crowdsourcing. In *Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services - MobiSys '13*, page 97. ACM Press, 2013.

[9] AppsFlyer. 360° Mobile Attribution. https://www.appsflyer.com/product/mobile-attribution-for-user-acquisition/, 2021.

[10] AppsFlyer. Android SDK integration for developers. https://support.appsflyer.com/hc/en-us/articles/207032126-Android-SDK-integration-for-developers#integration, 2021.

[11] Aurora Open Source Software. Warden : App management utility. https://gitlab.com/AuroraOSS/AppWarden, 2021.

[12] Rebecca Balebako, Abigail Marsh, Jialiu Lin, Jason Hong, and Lorrie Faith Cranor. The privacy and security behaviors of smartphone app developers. In *Proceedings 2014 Workshop on Usable Security*. Internet Society, 2014.

[13] Solon Barocas and Helen Nissenbaum. On notice: The trouble with notice and consent. In *Proceedings of the engaging data forum: The first international forum on the application and management of personal electronic information*, 2009.

[14] Elettra Bietti. Consent as a Free Pass: Platform Power and the Limits of the Informational Turn. *Pace Law Review*, page 60, 2020.

[15] Reuben Binns, Ulrik Lyngs, Max Van Kleek, Jun Zhao, Timothy Libert, and Nigel Shadbolt. Third party tracking in the mobile ecosystem. In *Proceedings of the 10th ACM Conference on Web Science - WebSci '18*, pages 23–31. ACM Press, 2018.

[16] Reuben Binns, Jun Zhao, Max Van Kleek, and Nigel Shadbolt. Measuring third-party tracker power across web and mobile. *ACM Transactions on Internet Technology*, 18(4):1–22, 2018.

[17] Marcel Bokhorst. NetGuard. https://github.com/M66B/NetGuard, 2021.

[18] Marcel Bokhorst. XPrivacyLua. https://lua.xprivacy.eu/, 2021.

[19] Bundeskartellamt. B6-22/16 (facebook v bundeskartellamt).

[20] Kai Chen, Xueqiang Wang, Yi Chen, Peng Wang, Yeonjoon Lee, XiaoFeng Wang, Bin Ma, Aohui Wang, Yingjun Zhang, and Wei Zou. Following Devil's Footprints: Cross-Platform Analysis of Potentially Harmful Libraries on Android and iOS. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 357–376. IEEE, 2016.

[21] Saksham Chitkara, Nishad Gothoskar, Suhas Harish, Jason I. Hong, and Yuvraj Agarwal. Does this app really need my location?: Context-aware privacy management for smartphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–22, 2017.

[22] Commission Nationale de l'Informatique et des Libertés. Décision n° MED 2018-042 du 30 octobre 2018 mettant en demeure la société X. https://www.legifrance.gouv.fr/cnil/id/CNILTEXT000037594451/, 2018.

[23] Competition and Markets Authority. Online platforms and digital advertising. https://assets.publishing.service.gov.uk/media/5efc57ed3a6f4023d242ed56/Final_report_1_July_2020_.pdf.

[24] Court of Justice of the European Union. Tietosuojavaltuutettu. http://curia.europa.eu/juris/

document/document.jsf?docid=203822&doclang=EN, 2018.

[25] Court of Justice of the European Union. Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein v Wirtschaftsakademie Schleswig-Holstein GmbH. http://curia.europa.eu/juris/liste.jsf?num=C-210/16, 2018.

[26] Court of Justice of the European Union. Fashion ID GmbH & Co. KG v Verbraucherzentrale NRW e. V. http://curia.europa.eu/juris/liste.jsf?num=C-40/17, 2019.

[27] Court of Justice of the European Union. Orange România SA v Autoritatea Naţională de Supraveghere a Prelucrării. http://curia.europa.eu/juris/document/document.jsf?docid=233544&doclang=EN, 2020.

[28] Datenschutzkonferenz. Orientierungshilfe der Aufsichtsbehörden für Anbieter von Telemedien.

[29] Disconnect.me and Mozilla. Firefox Blocklist. https://github.com/mozilla-services/shavar-prod-lists.

[30] Benjamin G Edelman and Damien Geradin. Android and competition law: Exploring and assessing google's practices in mobile. *European Competition Journal*, 2016.

[31] Manuel Egele, Christopher Kruegel, Engin Kirda, and Giovanni Vigna. PiOS: Detecting Privacy Leaks in iOS Applications. In *Proceedings of NDSS 2018*, 2011.

[32] Anirudh Ekambaranathan, Jun Zhao, and Max Van Kleek. "Money makes the world go around": Identifying barriers to better privacy in children's apps from developers' perspectives. In *Conference on Human Factors in Computing Systems (CHI '21)*, pages 1–24. ACM Press, 2021.

[33] William Enck, Peter Gilbert, Byung-Gon Chun, Landon P. Cox, Jaeyeon Jung, Patrick McDaniel, and Anmol N. Sheth. TaintDroid: An Information-flow Tracking System for Realtime Privacy Monitoring on Smartphones. In *Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation*, OSDI'10, pages 393–407, 2010.

[34] EU Article 29 Data Protection Working Party. Opinion 1/2010 on the concepts of "controller" and "processor". https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2010/wp169_en.pdf, 2010.

[35] EU Article 29 Data Protection Working Party. Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC. https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf, 2014.

[36] EU Article 29 Data Protection Working Party. Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679. https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=611236, 2017.

[37] European Data Protection Board. Guidelines 07/2020 on the concepts of controller and processor in the GDPR. https://edpb.europa.eu/our-work-tools/documents/public-consultations/2020/guidelines-072020-concepts-controller-and_en, 2020.

[38] European Parliament and Council. Regulation 2016/679 (General Data Protection Regulation). http://data.europa.eu/eli/reg/2016/679/oj, 4 2016.

[39] Exodus. Statistics. https://reports.exodus-privacy.eu.org/en/trackers/stats/.

[40] Ronan Ó Fathaigh. Mobile privacy and business-to-platform dependencies: An analysis of SEC disclosures. *Journal of Business*, page 58, 2018.

[41] Eric Ferrari-Herrmann. Trapped in Google's safety net: what modders need to know. , 2021.

[42] Rudolph Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948.

[43] Google. Requesting Consent from European Users. https://developers.google.com/admob/android/eu-consent.

[44] Google. Android 7.0 for Developers. https://developer.android.com/about/versions/nougat/android-7.0#default_trusted_ca, 2016.

[45] Google. Device and network abuse. https://support.google.com/googleplay/android-developer/answer/9888379, 2021.

[46] Google. Get started with AdMob in your Android project. https://firebase.google.com/docs/admob/android/quick-start, 2021.

[47] Google. User Data. https://support.google.com/googleplay/android-developer/answer/10144311, 2021.

[48] Daniel Greene and Katie Shilton. Platform privacies: Governance, collaboration, and the different meanings of "privacy" in iOS and android development. *New Media & Society*, 20(4):1640–1657, 2018.

[49] Irit Hadar, Tomer Hasson, Oshrat Ayalon, Eran Toch, Michael Birnhack, Sofia Sherman, and Arod Balissa. Privacy by designers: software developers' privacy mindset. *Empirical Software Engineering*, 23(1):259–289, 2018.

[50] Catherine Han, Irwin Reyes, Amit Elazari, Joel Reardon, Alvaro Feal, Kenneth A. Bamberger, Serge Egelman, and Narseo Vallina-Rodriguez. Do you get what you pay for? comparing the privacy behaviors of free vs. paid apps. In *The Workshop on Technology and Consumer Protection (ConPro '19)*, 2019.

[51] Jin Han, Qiang Yan, Debin Gao, Jianying Zhou, and Robert H Deng. Comparing Mobile Privacy Protection through Cross-Platform Applications. In *Proceedings 2013 Network and Distributed System Security Symposium*, page 16. Internet Society, 2013.

[52] Information Commissioner's Office. How do we apply legitimate interests in practice? . https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/legitimate-interests/how-do-we-apply-legitimate-interests-in-practice/, 2021.

[53] Information Commissioner's Office. How do we comply with the cookie rules? https://ico.org.uk/for-organisations/guide-to-pecr/guidance-on-the-use-of-cookies-and-similar-technologies/how-do-we-comply-with-the-cookie-rules/, 2021.

[54] Information Commissioner's Office. What are the rules on cookies and similar technologies? https://ico.org.uk/for-organisations/guide-to-pecr/guidance-on-the-use-of-cookies-and-similar-technologies/what-are-the-rules-on-cookies-and-similar-technologies/#rules10, 2021.

[55] Inmobi. Android Guidelines: Getting Started with Android SDK Integration. https://support.inmobi.com/monetize/android-guidelines/, 2021.

[56] Caitlin Kelleher and Michelle Ichinco. Towards a model of api learning. In *2019 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 163–168. IEEE, 2019.

[57] Julian Andres Klode. DNS-Based Host Blocking for Android. https://github.com/julian-klode/dns66, 2021.

[58] Joseph Lawrance, Christopher Bogart, Margaret Burnett, Rachel Bellamy, Kyle Rector, and Scott D Fleming. How programmers debug, revisited: An information foraging theory perspective. *IEEE Transactions on Software Engineering*, 39(2):197–215, 2010.

[59] Anh Le, Janus Varmarken, Simon Langhoff, Anastasia Shuba, Minas Gjoka, and Athina Markopoulou. Antmonitor: A system for monitoring from mobile devices. In *Proceedings of the 2015 ACM SIGCOMM Workshop on Crowdsourcing and Crowdsharing of Big (Internet) Data*, C2B(1)D '15, pages 15–20, 8 2015.

[60] Célestin Matte, Nataliia Bielova, and Cristiana Santos. Do cookie banners respect my choice? measuring legal compliance of banners from IAB europe's transparency and consent framework. *2020 IEEE Symposium on Security and Privacy (SP)*, 2019.

[61] Aleecia M McDonald and Lorrie Faith Cranor. The Cost of Reading Privacy Policies. *I/S: A Journal of Law and Policy for the Information Society*, page 26, 2008.

[62] Abraham H Mhaidli, Yixin Zou, and Florian Schaub. "We Can't Live Without Them!" App Developers' Adoption of Ad Networks and Their Considerations of Consumer Risks. *Proceedings of the Fifteenth Symposium on Usable Privacy and Security*, page 21, 2019.

[63] microg. Implementation Status. https://github.com/microg/GmsCore/wiki/Implementation-Status, 2020.

[64] Patricia A. Norberg, Daniel R. Horne, and David A. Horne. The Privacy Paradox: Personal Information Disclosure Intentions versus Behaviors. *Journal of Consumer Affairs*, 41(1):100–126, 2017.

[65] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[66] Ehimare Okoyomon, Nikita Samarin, Primal Wijesekera, Amit Elazari, Narseo Vallina-Rodriguez, Irwin Reyes, Álvaro Feal, and Serge Egelman. On The Ridiculousness of Notice and Consent: Contradictions in App Privacy Policies. *The Workshop on Technology and Consumer Protection (ConPro '19)*, 2019.

[67] Emmanouil Papadogiannakis, Panagiotis Papadopoulos, Nicolas Kourtellis, and Evangelos P. Markatos. User Tracking in the Post-cookie Era: How Websites Bypass GDPR Consent to Track Users. *WWW'2021*, 2021.

[68] Abbas Razaghpanah, Rishab Nithyanand, Narseo Vallina-Rodriguez, Srikanth Sundaresan, Mark Allman, Christian Kreibich, and Phillipa Gill. Apps, trackers, privacy, and regulators: A global study of the mobile tracking ecosystem. In *Proceedings of NDSS 2018*, 2 2018.

[69] Joel R. Reidenberg, Jaspreet Bhatia, Travis D. Breaux, and Thomas B. Norton. Ambiguity in Privacy Policies and the Impact of Regulation. *The Journal of Legal Studies*, 45(S2):S163–S190, 2016.

[70] Jingjing Ren, Ashwin Rao, Martina Lindorfer, Arnaud Legout, and David Choffnes. ReCon: Revealing and Controlling PII Leaks in Mobile Network Traffic. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services - MobiSys '16*, pages 361–374. ACM Press, 2016.

[71] Irwin Reyes, Primal Wijesekera, Joel Reardon, Amit Elazari Bar On, Abbas Razaghpanah, Narseo Vallina-Rodriguez, and Serge Egelman. "Won't Somebody Think of the Children?" Examining COPPA Compliance at Scale. *Proceedings on Privacy Enhancing Technologies*, 2018(3):63–83, jun 2018.

[72] Irina Shklovski, Scott D. Mainwaring, Halla Hrund Skúladóttir, and Höskuldur Borgthorsson. Leakiness and creepiness in app space: Perceptions of privacy and mobile app use. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI '14*, pages 2347–2356. ACM Press, 2014.

[73] Anastasia Shuba, Athina Markopoulou, and Zubair Shafiq. Nomoads: Effective and efficient cross-app mobile ad-blocking. In *Proceedings on Privacy Enhancing Technologies 2018*, pages 125–140, 10 2018.

[74] Daniel J. Solove. Privacy Self-Management and the Consent Dilemma. *Harvard Law Review*, 2012.

[75] Daniel J Solove. The myth of the privacy paradox. *Available at SSRN*, 2020.

[76] Yihang Song and Urs Hengartner. Privacyguard: A vpn-based platform to detect information leakage on android devices. In *Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices*, SPSM '15, pages 15–26, 2015.

[77] Juan Carlos Torres. Google SafetyNet update might be the end for Android rooting, custom ROMs. https://www.slashgear.com/google-safetynet-update-might-be-the-end-for-android-rooting-custom-roms-30627121/, 2020.

[78] UK Competition and Markets Authority. Online platforms and digital advertising market study final report. https://assets.publishing.service.gov.uk/media/5efc57ed3a6f4023d242ed56/Final_report_1_July_2020_.pdf, 2020.

[79] United States Congress. Children's Online Privacy Protection Act. https://www.ftc.gov/system/files/2012-31341.pdf, 1998.

[80] Max Van Kleek, Ilaria Liccardi, Reuben Binns, Jun Zhao, Daniel J. Weitzner, and Nigel Shadbolt. Better the Devil You Know: Exposing the Data Sharing Practices of Smartphone Apps. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, pages 5208–5220. ACM Press, 2017.

[81] Nicolas Viennot, Edward Garcia, and Jason Nieh. A measurement study of google play. In *The 2014 ACM International Conference on Measurement and Modeling of Computer Systems*, SIGMETRICS '14, pages 221–233, 2014.

[82] Ari Ezra Waldman. Cognitive biases, dark patterns, and the 'privacy paradox'. *Current opinion in psychology*, 31:105–109, 2020.

[83] Haoyu Wang, Zhe Liu, Jingyue Liang, Narseo Vallina-Rodriguez, Yao Guo, Li Li, Juan Tapiador, Jingcun Cao, and Guoai Xu. Beyond google play: A large-scale comparative study of chinese android app markets. In *Proceedings of the Internet Measurement Conference 2018*, IMC '18, pages 293–307, 2018.

[84] XDA Developers. Google Play Services will soon delete your advertising ID when you opt out of ad personalization. https://www.xda-developers.com/google-play-services-delete-ad-id-opt-out-personalization/, 2021.

[85] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N Cameron Russell, and Norman Sadeh. Maps: Scaling privacy compliance analysis to a million apps. *Privacy Enhancing Technologies Symposium 2019*, 72, 6 2019.

"Whether it's moral is a whole other story": Consumer perspectives on privacy regulations and corporate data practices

Leah Zhang-Kennedy
University of Waterloo

Sonia Chiasson
Carleton University

Abstract

Privacy laws govern the collection, use, and disclosure of personal information by businesses. Through an online survey with 300 participants and a follow-up interview with 32 participants, we investigate Canadians' awareness of their privacy rights and how businesses manage their personal information. Further, we explore how Canadians respond to hypothetical privacy violations using ten scenarios adapted from real cases. Our participants are generally aware of having privacy rights but have insufficient knowledge and resources to exercise those rights properly. Participants did not necessarily equate compliance with the law as sufficient for ethical conduct. Through our analysis, we identified a "moral code" that consumers rely on to assess privacy violations based on the core moral values of trust, transparency, control, and access.

1 Introduction

Despite rapid technological change, the Canadian regulatory landscape under the Fair Information Practices Principles framework that has governed consumer privacy has remained largely unchanged for the last 50 years [8]. From this framework, Canada has devised ten Fair Information Principles (FIPs) under Canadian privacy law known as the Personal Information Protection and Electronic Documents Act (PIPEDA). The core principles of PIPEDA are: 1) accountability, 2) identifying purposes, 3) consent, 4) Limiting Collection, 5) Limiting Use, Disclosure, and Retention, 6) Accuracy, 7) Safeguards, 8) Openness, 9) Individual Access, and 10) Challenging Compliance (described in Appendix A4.1).

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

We investigated Canadians' perspectives on their privacy rights and corporate data practices relating to their digital data through a survey with 300 Canadian residents and followed-up with 32 interviews. The studies explored general privacy perceptions and self-reported knowledge of businesses' data collection and usage practices towards consumer data. Participants described their understanding of their own privacy rights and their interpretations of ten scenarios describing corporate data privacy practices adapted from real privacy cases published online [24] by the Office of the Privacy Commissioner of Canada (OPC) to guide compliance with PIPEDA.

Our work makes two main contributions. First, we expand the literature on individuals' privacy perspectives and understanding about corporate data practices. Participants perceived significant challenges to consumer privacy protection: a lack of awareness, difficulty enforcing privacy laws, rapid technological change, and safeguarding against hackers. They were largely unaware of the PIPEDA FIPs and unsure how they applied to the provided scenarios. The interviews uncovered that participants relied on an informal "moral code" to judge privacy violations. This code was derived from personal values of trust, transparency, control, and access. Participants wanted businesses to follow this moral code even when it exceeded legal requirements.

Second, our mixed-study methodology enables a better understanding of users' reasoning and interpretation of the situation when faced with privacy violations. Participants identified various barriers that prevented them from raising privacy concerns with businesses or regulatory bodies, even though most feel it was primarily the consumer's responsibility to report such privacy violations. We observed ambivalence from participants, as they felt that individuals were largely powerless when faced with corporate privacy violations, regardless of whether these violated regulations or their own moral code. Our work increases awareness of end-user perspectives among stakeholders and supports calls for change. It can also inform educational efforts and may prompt privacy-supportive systems to help users manage their privacy in this context.

2 Background

Personal information on digital platforms could be exposed to other individuals, but may also be collected, used and shared with institutions [7]. Social privacy involves privacy situations with other individuals, whereas institutional privacy concerns users' relationships with organizations who collect, use, and share their personal data to provide online services [30]. Past research shows that users tend to focus their privacy concerns on the appropriateness of shared information in a social context and neglect institutional privacy risks [30]. A study that focuses on social media users [32] found that privacy is understood universally as a matter of controlling one's own data relating to personal autonomy and that concerns and engagement in protective tactics centres on being personally affected by privacy violations. Many consumers feel ill-informed about how their data is collected and used [27, 28, 31].

Even though users recognize a need to protect their private data, many feel they have little control over their own data [32]. Mayer's Integrative Model of Organizational Trust [22] posits that trust is the perception of an organization's ability, benevolence, and integrity. The perceived effectiveness of privacy legislation and the trust users have towards organizations could affect their perceived effectiveness of privacy policies, perceived benefits of information disclosure, and their assessments of online privacy risks [41].

Privacy has been described in several distinct but related theories. Privacy could be described as the right to be left alone [34]. Yet, this does not capture the relationship between consumers and corporate organizations. When consumer behaviour is observed in context of privacy, a paradox is often observed [14, 19]. Reasons for the privacy paradox commonly describe consumers' lack of awareness [12, 27] and the notion of privacy as a commodity: trading personal information for convenience, goods, and services [3, 5]. Other theories define privacy as a state subjective to individuals' perceptions and beliefs [34]. Altman [6] defines privacy as the "the selective control of access to the self or to one's group." Similarly, Westin [42] described privacy as the perceived control an individual has over the collection and use of personal information. Solove [35] argues that privacy has many different meanings serving various functions in different contexts. The notion of *contextual integrity* considers the flow of information about individuals that are related to the context and could be violated when the informational norms associated with a given situation are breached [23]. Though context-dependent privacy research (e.g., [18, 39]) enables inferences about privacy decisions, further research is needed to identify conditions that lead to disclosure decisions [18].

In a recent position paper, Abdul-Ghani [1] positioned the extent to which consumers are aware of the data collection mechanism used by organizations and the tools available to consumers to protect their personal information as an ethical issue that could impact modern digital marketing practice. For example, institutional privacy assurances such as privacy policies can help to reduce individual privacy concerns [44]. The problem is that users seldom read privacy policies before agreeing to the terms and conditions because policies are long and difficult to understand [4]. Palmatier and Martin [26] recommended several ways for organizations to act ethically regarding the collection, use, storage, and dissemination of consumer data, including minimizing data collection, more transparency and control, and protecting data from data breaches, and regular audits of organizational privacy practices. Of course, other competing priorities for organizations may render these options less desirable from their perspective than their current practices, especially if current practices technically comply with the relevant regulations.

Some researchers [25] have proposed privacy as a dynamic, dialectic process, where privacy regulation is under continuous negotiation and management conditioned by one's own expectations and experiences. However, existing research on users' understanding of privacy rights shows that although many like the concepts of having privacy rights, users generally do not know what their rights are [17]. Furthermore, since users seldom read privacy policies, their expectations regarding corporate data practices are often mismatched against the actual data practices, leading to unintended sharing of personal information online [29].

3 Methodology

Our mixed study methodology was cleared by our university's research ethics board and consisted of a survey and follow-up one-on-one interviews with a subset of participants. We consulted a law and privacy expert during the development of the survey and pilot tested with lab members. We collected 300 survey responses from Canadian residents using Prolific[1] for recruitment. Participants (148 self-identifying as male, 149 as female, and 3 as non-binary) were compensated $3.40 CAD for completing a Qualtrics[2] questionnaire, which took on average 14.2 minutes to complete ($SD = 8$ minutes). Table 1 summarizes our participants' demographics. We had more participants from the province of Ontario and in the 20s to 30s age range with lower levels of education and income compared to the most recent Statistics Canada census data [37]. Note that we did not exclude participants from Quebec (QC), but the province is primarily French-speaking. Thus, we believe that this impacted their interest in our English-language survey on Prolific. Using Westin's privacy clusters, 8% of participants are marginally concerned (i.e., low privacy concern), 75% are pragmatists (i.e., medium privacy concern), and 16% are fundamentalists (i.e., high privacy concern). We found reasonable agreement between our user clusters compared to

[1] https://www.prolific.co
[2] https://www.qualtrics.com

past studies [2, 10, 13, 33].

From the survey sample, we pseudo-randomly invited (i.e., ensuring broad coverage of demographics) 32 interested participants to a virtual follow-up interview that lasted on average 39 minutes ($SD = 11$ minutes). Each interview participant was compensated $20.70 CAD. The participants' identities were anonymized with a code name (e.g., P1, P32)

3.1 Survey

The survey (see Appendix A) contained Likert-scale and multiple-choice questions with a "Prefer not to answer" option for all questions. The survey is divided into four sections, with the first section containing demographic information. The second section included Westin's privacy index questions. The third section focused on self-reported knowledge of Canadian privacy regulations, privacy rights and protection, how businesses collect, use, and share personal information, and perceptions of smart technology's impact on privacy.

In the fourth section, each participant was randomly assigned to five out of ten privacy vignettes created from real privacy complaints against organizations, investigated by the Office of the Privacy Commissioner of Canada (OPC). Randomizing five of ten vignettes enabled us to explore a broader range of data privacy scenarios without overburdening the participants. Each case's conclusions are based on factual analysis through court decisions and OPC findings, which provide reasonable guidelines for whether the organization's actions were in compliance or violation of a provision of PIPEDA. We selected ten cases with clear outcomes, covering a range of FIPs, and that are likely to occur in everyday life from twenty candidate cases. For brevity, we summarized the scenarios in Table 2. More information about the selected cases is available in Appendix A4.2.

Each vignette was displayed one at a time and accompanied by three five-point Likert-scale questions (Strong agree to Strongly disagree); the questions are: 1) I think scenarios like this are likely to happen; 2) I would be concerned about my privacy in this scenario; 3) I think the business acted appropriately in a lawful manner based on the situation described. Lastly, we asked participants to select "Which of the privacy principles do you think apply in this situation" from a checklist. To ensure a baseline understanding of the ten FIPs, we displayed the OPC's official descriptions of the principles for each scenario.

3.2 Interview

Approximately one-third ($n = 111$) of the survey participants volunteered to be contacted via email for a follow-up interview. We sent these participants a screening questionnaire containing the interview consent form; 79 participants responded, and 53 agreed to schedule an interview. In the final stage, 32 participants completed the interview via video conferencing.

The semi-structured interview consisted of two parts (see Appendix A2). In the first part, we asked general questions regarding personal information and how Canadian privacy laws protect consumer privacy. We then asked participants to explain whether they think companies and existing laws provide adequate protection and what other protections should exist. We inquired about whose responsibility it is to report privacy concerns. If the participants had a previous privacy concern or complaint against a business, they recounted the incident. Lastly, participants shared their thoughts about the biggest challenges facing consumer privacy protection.

In the second part, the participants clarified and elaborated on their responses to their previously completed vignette scenarios in the survey. We were particularly interested in their opinions about whether the business had acted appropriately under the law and whether they would be concerned about their privacy if faced with the scenario. If relevant, they were asked to share a similar situation they experienced. We re-read the scenarios from the survey before participants responded and encouraged them to thoughtfully discuss their responses to the scenarios. The interviews were audio-recorded, then transcribed using Trint[3] speech-to-text software and manually checked for accuracy.

3.2.1 Grounded Theory analysis

We chose Grounded Theory methodology [11] to analyze the interview data to form an explanatory theory about how consumers assess privacy violations in the collection, use, and disclosure of their personal information. In the first iteration, the lead researcher read all transcripts to gain an overall understanding, then coded all transcripts point-by-point in Atlas.ti[4] qualitative analysis software and developed 106 descriptive codes. Through Axial coding, we developed a codebook by looking for patterns and connections within the codes, and generated 13 groups. Figure 1 shows a sample of the codes grouped into higher-level concepts.

A research assistant used the developed codebook to conduct a second independent analysis of 10 out of 32 interview transcripts). We used Krippendorff's alpha coefficient [20] to measure the agreement of the two coders because it is sensitive to small samples, whereas Cohen's kappa assumes an infinite sample size [21]. Krippendorff [20] suggests $\alpha \geq 0.667$ as the minimum acceptable value. Our test showed moderate agreement between the two researchers' analyses, $\alpha = 0.741$. The two researchers met and resolved the coding variability by explaining their rationale for the analysis and discussed until they reached a mutual agreement. The lead researcher then re-coded the remaining interview based on the agreed

[3] https://trint.com
[4] https://atlasti.com

Province and Territory				Gender				Age Group				Level of Education				Income			
	Survey	StatCan	Interview		Survey	StatCan	Interview		Survey	StatCan	Interview		Survey	StatCan	Interview		Survey	StatCan	Interview
ON	55%	(38%)	59%	Male	49%	(49%)	50%	18 to 19 years	4%	(N/A)	0%	No high school	1%	(12%)	3%	<$15k	6%	(21%)	0%
BC	16%	(13%)	19%	Female	50%	(51%)	50%	20 to 29 years	42%	(13%)	22%	High school	18%	(24%)	6%	$15k-$34k	16%	(30%)	16%
AB	12%	(12%)	13%	Non-binary	1%	(N/A)	0%	30 to 39 years	36%	(14%)	50%	College	14%	(22%)	19%	$35k-$74k	34%	(33%)	31%
NS	6%	(3%)	6%					40 to 49 years	10%	(13%)	13%	Bachelors or higher	64%	(29%)	66%	$75k-$149k	31%	(14%)	38%
MB	4%	(4%)	3%					50 to 59 years	4%	(15.0%)	9%	Other Professional	2%	(11%)	6%	$150k-$199k	4%	(2%)	6%
SK	3%	(4%)	0%					60+ years	3%	(23%)	6%	No answer	0.3%	(N/A)	0%	>$200k	2%	(2%)	6%
NL	1%	(2%)	0%													No answer	6%	(N/A)	3%
PE	1%	(0.4%)	0%																
NB	0.7%	(2%)	0%																
QC	0%	(23%)	0%																
YT	0.3%	(0.1%)	0%																
NT	0%	(0.1%)	0%																
NU	0%	(0.1%)	0%																

Table 1: Participant demographic information for the survey and interview study. Survey demographics are compared to national averages from Statistics Canada's most recent census data (in brackets).

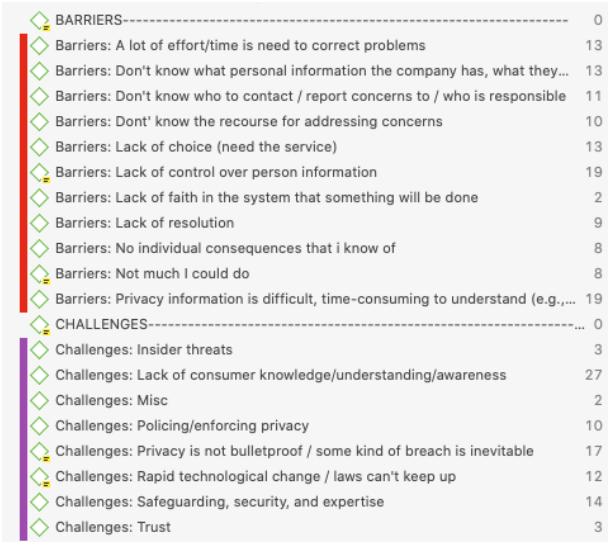


Figure 1: A subset of codes used in the open coding process in Atlas.ti. The codes are grouped into related concepts based on the axial coding process in the format "Concept: Code".

analysis. Lastly, we used Selective Coding to integrate results into a theory unifying core themes and grounded in the data.

4 Survey Results

The majority of participants owned at least two types of internet-connected devices. Desktop, laptops, and mobile phones are the most common (99%), followed by tablets (70%), gaming consoles (68%), and smart media devices (65%). Less than half owned home assistants (42%), wearables (37%), and smart appliances (37%). Some have a car with a smart system (19%) and home security systems (12%); few have internet-connected toys, monitors, and trackers (8%), and medical health monitors (3%).

4.1 Technology's impact on privacy

Only 36% rated their knowledge of how these technologies affected their privacy as good or very good. As summarized in Figure 2, participants felt they had poor knowledge about how new technology like home assistants, smart devices, and smart cars, and other connected devices affect their privacy. Even though they reported having highest knowledge about how desktops and laptops, followed by tablets, home security systems, and mobile phones affect their privacy, participants were not very confident about these either.

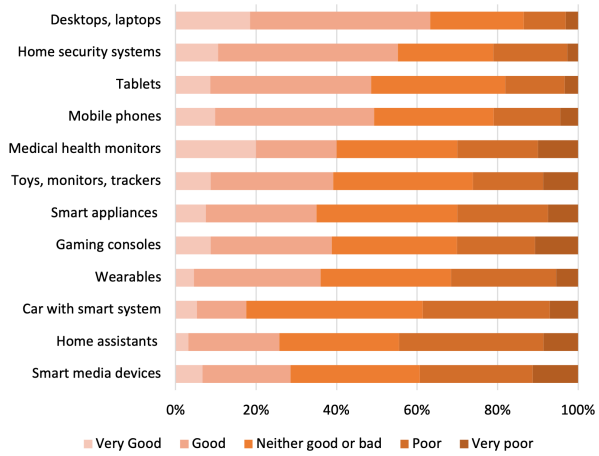


Figure 2: Self-reported knowledge of how to protect personal information across a variety of devices.

4.2 Information management practices

As summarized in Figure 3, participants reported being diligent in applying some information privacy management practices while neglecting others. We note that these are self-reported behaviours. Therefore, they may not fully reflect actual behaviours. Participants believe themselves to be most diligent in avoiding sharing their Social Insurance Number (SIN), exercising safe password practices, downloading files from reputable sources, and installing the latest software updates. They reported being less attentive about using encryption and disabling Wi-Fi and Bluetooth when not in use and when moving through public spaces. More than half would withhold sharing optional information, but less than half think about why their data is needed, who will use it, and how it

Scenario	Description	Compliant	Principles
S1-outsourcing-abroad	Your email provider notifies you that your email subscription will be outsourced to the US. You will be asked to accept or decline the new services upon login to your new account.	Yes	Accountability; Consent
S2-GPS-tracking	Your telecommunications employer notifies you that they will begin tracking your location via Global Positioning System (GPS) on company vehicles to manage workforce productivity, safety, and company assets.	Yes	Identifying Purposes; Consent; Limiting Collection; Limiting Use; Safeguards; Openness
S3-opt-out-consent	Your cellular provider notifies you by mail that the company intends to use customers' personal information for secondary marketing purposes. You could have your name removed from the marketing list by contacting the company; otherwise, it will assume your consent.	Yes	Consent
S4-over-collection	You are asked for your personal identification information (Utility bill and driver's licence) for the purpose of verifying your identify for receiving a free $10 gift card.	No	Accountability; Consent; Limiting Collection; Openness
S5-amending-consent	You receive a notice from your bank that its changing their policy to use your personal information for the secondary purpose of marketing. The notice outlines who would have access to customers' personal information and how to withdraw your consent.	No	Consent
S6-identify-theft	Your personal information was used by a fraudster to open a credit card account using your personal information, and your bank assumed the financial loss for the account balance	No	Accuracy; Safeguards
S7-safeguarding-data	A connected toy manufacturer, of which you are a customer, notifies you that they are improving security after a data breach resulting in the potential compromise of you and your child's personal information.	No	Safeguards
S8-openness-of-collection	You are asked to create a User ID and provide your credit card information to access online services from a well-known technology company to download a free app. Instructions for downloading without providing the information is posted in the website's support section.	No	Identifying Purposes; Limiting Collection; Openness
S9-accessing-password	Your request to directly access your login-related information (date, time, and IP address) from a web-based company after suspicious password reset is denied based on the explanation that only law enforcement can have access, not clients.	No	Accountability; Safeguards; Individual Access; Challenging Compliance
S10-challenging-exceptions	Your physician refuses to provide your insurance company his personal notes after your medical examination because he claims it is not part of your official medical record.	No	Individual Access

Table 2: Scenario descriptions (condensed version) and the relevant privacy principles. The OPC ruled the first three scenarios compliant with PIPEDA and the rest in violation.

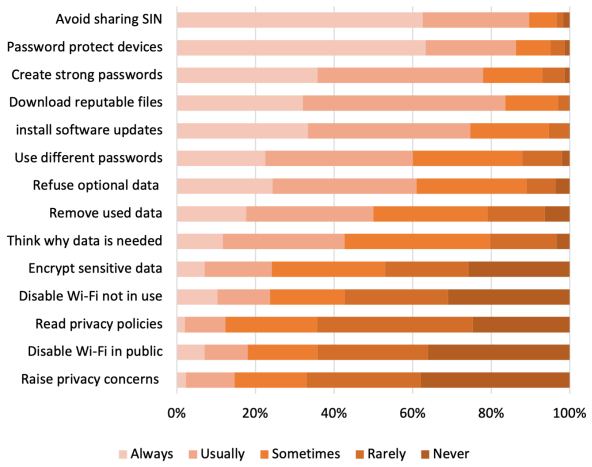


Figure 3: Self-reported information privacy behaviour.

would be used before providing it online. Unsurprisingly, many participants do not read privacy policies. Only half said they would remove their personal information when they no longer need the product or service. Even though 67% of participants indicated that they are concerned about their personal information held by companies, few said they would raise a privacy concern if companies mishandled their personal information.

4.3 Awareness of privacy rights

Only a third of participants indicated that they have good knowledge of their privacy rights (29%) and how to protect those rights (37%). This overall low level of knowledge is reflected in the low awareness of how businesses manage their personal information. Most participants are aware of the information management practices for only some or none of the services and products that they use. We used Friedman's Analysis of Variance to determine how their awareness differed across the eight types of data management practices (Figure 4). We found an overall statistically significant difference between perceived awareness of different practices ($\chi^2(8) = 559.987, p < .0005$). Pairwise comparisons with a Bonferroni correction for multiple comparisons revealed a statistically greater perceived awareness for *what is collected*, *why it is collected*, and *how it is collected* compared to the other practices.

4.4 Applying FIPs

We first asked the participants whether they felt the scenario were likely to happen in real life to ensure that our selection of scenarios was relatable. Over 75% of participants agree that the majority of scenarios (S2-S3, S5-S8) are likely to happen. Over half of participants agreed that the remaining

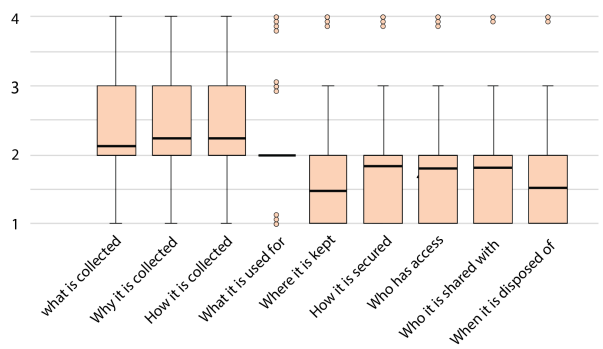


Figure 4: Perceived awareness of ways businesses manage personal information for services and products currently used; Likert scale responses: 1 = none, 2 = some, 3 = most, 4 = all

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Accountability	58%	54%	48%	51%	47%	82%	81%	51%	60%	44%
Identifying Purposes	35%	65%	58%	81%	65%	44%	32%	55%	33%	40%
Consent	74%	55%	81%	66%	77%	39%	24%	52%	46%	47%
Limiting Collection	34%	49%	42%	64%	42%	23%	41%	54%	32%	41%
Limiting Use	43%	47%	48%	62%	46%	26%	36%	42%	44%	49%
Accuracy	15%	34%	18%	19%	18%	56%	14%	19%	19%	25%
Safeguards	41%	42%	26%	47%	43%	79%	81%	36%	61%	27%
Openness	58%	55%	56%	47%	57%	31%	47%	55%	32%	50%
Individual Access	45%	40%	35%	33%	37%	45%	28%	33%	53%	59%
Challenging Compliance	40%	32%	25%	32%	38%	38%	26%	31%	46%	43%
Don't know	11%	7%	4%	3%	9%	3%	5%	11%	9%	13%

Figure 5: Percentage of participants who applied the FIPs to each scenario (S1-S10). The blue scale represents the principles used by the OPC in its official case interpretations. Darker cells represent a higher percentage.

three scenarios were likely (55% to 68%). After reading each scenario, the participants selected the FIPs they thought would apply to the situation. Figure 5 summarizes the percentage of participants who selected each principle per scenario compared to the OPC interpretations. Our participants generally over-applied the principles to privacy situations. This may be because they have insufficient understanding of the principles, or insufficient detail to appreciate the scenarios' nuances fully. We would not necessarily assume that a layperson would have perfect alignment with the OPC, but these responses give a general sense of their interpretations. Laypersons' misapplication of the FIPs (compared to regulators) suggests that consumers (i) have misconceptions of their privacy rights and (ii) have low efficacy to hold organizations accountable for privacy violations.

5 Interview Results

This section reports our qualitative findings regarding participants' understanding of personal information and how Canadian privacy laws protect consumer privacy. We recorded the frequencies during data analysis to help with identifying trends, but deliberately avoided reporting numbers in the paper, as is recommended for inductive approaches like Grounded Theory [16]. Instead, we use descriptive language (e.g., most, some, few, none) where appropriate. Supplementary interview results that are not central to our research question are included in Appendix C.

5.1 Canadian privacy protection

Most participants admitted to being "unaware of what privacy laws are and what is required of companies" or unclear about the "specifics" of what the law says, but "do know that there are laws in place." P5 explained,

> *I don't know the letter of the law and what the laws specifically are, but my gut feeling is that privacy or personal information probably isn't super well protected... because so much information gets put onto the Internet... I don't doubt that there are laws in place that try to protect that as much as possible. I just see it kind of as an inevitable thing that information will leak out one way or another. (P5)*

5.1.1 Effectiveness of existing privacy laws

Participants' overall consensus is that the law offers "weak", "ineffective", and "unregulated" protection. Some preferred other "hardcore" international regulatory bodies like the European GDPR. P5 explained, *"I think of what I read, the [Canadian] laws... sound really good. I just doubt whether they are put into place in a way that actually protects information."* P7 also believed not enough is being done: *"I don't know if that's from a lack of [laws], you know, how the requirements are written or if it's on the side of them enforcing the rules. It feels like not enough is being done. I don't know exactly why or where that is."*

Some believe businesses do their best to protect consumers' information, but it is not *"bulletproof"* (P2). *"Generally"*, said P20, *"I'm assuming [companies] have security systems in place... because of the idea if something was breached and something leaked out, it would be bad publicity for that company. So I think they're trying to do as best they can."*

Another group believed businesses have no incentives to protect consumer privacy because *"information and data are hugely profitable for companies"* (P1). Because information is valuable, it is *"in their financial best interest to obfuscate what they're collecting"* (P7).

Others simply "don't know" whether companies provide reasonable protection. P4 declared, *"I feel like I don't necessarily have enough understanding of how our information is being used... so I'm not sure I have an opinion on like what... because I don't really know what is a reasonable level of privacy... and what companies are doing right now."*

5.1.2 Responsibility for reporting privacy violations

The majority believed that the affected individuals (e.g., consumers) *"who felt that their privacy was violated"* (P5) are responsible for reporting concerns *"to the proper channels and take care of the problem"* (P2). Unfortunately, none of our participants could clearly identify the "proper channels" or who is accountable for businesses' compliance.

These participants internalized privacy violations as something that happens to them personally, and therefore they should be responsible for reporting. When asked about why consumers should report privacy concerns, P18 thought is it because *"[consumers] are the only ones that are concerned about our privacy... The companies are not going to bring it up... unless it involves a lot of money... and reaches the news."* Another believed the *"onus [is] on people to be to help themselves be informed about things... [otherwise] you're susceptible to being taken advantage of... and people using your information in an unethical way"* (P21).

The most common recourse is to report concerns directly to the business, but some also believed it is their responsibility to report to government agencies because *"the government wouldn't know unless you report them"* (P13). Many participants assumed the existence of a federal authority and government agency like a "privacy commissioner", "better business bureau", or "ombudsman", but none of the participants were aware of the process for reporting. *"I don't even know who to go to,"* said P23, *"I'm sure there's someone in government that's responsible for it... it seems like an owner's task to try to figure that out and lodge a complaint that probably will fall, if I'm being realistic, on deaf ears."*

A small number of participants believed that anyone who is aware of the privacy violation, like *"conscientious employees should whistle blow if they see something going on illegally"* (P7). Few mentioned that companies are responsible for bringing privacy violations forward to the consumer or a government agency because *"they have a legal and ethical responsibility"* (P11). Ultimately, participants thought that the responsibility *"falls on the consumer"* (P21) because *"your rights aren't really protected without you having to go out and do something on your own"* (P13).

5.1.3 Challenges for consumer privacy protection

Our participants identified four main challenges (C1–C4) for consumer privacy protection.

C1. Lack of awareness: Even though most participants believed consumers are responsible for reporting privacy violations, they also did not know how to address privacy concerns. For example, P11 had not raised any privacy concerns with companies because

> *I didn't know who I should raise that concern with. Should I bring it up with the company... send them an email... send it to some sort of privacy watchdog organization in the country? So I didn't raise a concern, but it wasn't because I didn't have a concern, I just didn't know what to do. (P11)*

Many participants felt that they lack awareness of the ramifications of information disclosure. P18 declared, *"the general population are... not aware of what not to provide to the companies,"* P11 elaborated,

> *As a Canadian consumer, there are so many things that, you know, I'm guilty of signing up for. I really have no clue what information [companies] have on me and how they're using it... It's not something that's clear to Canadians where to look, what they should advocate for. What's a reasonable expectation of information to give up? What's unreasonable? (P11)*

On the enforcement side, P31 admitted, *"I don't really know what the government does to ensure that information is being stored correctly and securely, or even collected lawfully. My perception has always been that it's the sort of thing that only gets dealt with when a problem comes up."*

The problem is that *"Canadians are not educated enough on the privacy laws that are available to protect their information,"* said P22. As one possible solution, our participants suggested more public awareness about the resources available. For example, *"finding out about the [privacy] commissioner of Canada that I didn't even know existed before now"* (P9).

C2. Enforcing privacy laws: Policing consumer privacy is a daunting task because of its "breadth" and "scope". *"Consumer privacy can be violated in so many different ways,"* explained P25, *"[it's] impossible to police every single application, every single website out there to see whether or not they're complying with whatever laws have been put in place."* Most participants recognized that many of the products and services they use operate in the United States or other countries: *"So many companies operate internationally that it's easy for some companies to sort of skirt around that... a company bases their servers in Thailand... whether a Canadian can enforce any sort of laws on that company is really questionable"* (P25). Therefore, privacy enforcement is viewed as *"an issue of scale with the incredible amount of data compared to... the limited resources of the government"* (P7). Enforcement is viewed as one of the *"biggest steps aside from the law itself"* (P14).

As a result of poor enforcement, our participants believe firmer laws with harsh penalties for non-compliance should be put in place, and they frequently used GDPR as an example of the type of enforcement they would *"like to see in Canada"* (P4). Highlighted protections included the right to permanently delete information and the right to refuse to provide information. If a company *"break the rules... they can get fined"* (P18). Some viewed these protections as *"what a company should have been doing already"*, but implementing the rules would *"force a lot more companies to adopt better privacy practices"* (P32).

C3. Rapid technological change: Our participants identified the reality that *"technology is evolving, and the law*

doesn't keep up" (P27). The "technology" mentioned include artificial intelligence, personal home assistants, and other "smart" devices. *"we're not even sure how to legislate for [these technologies]"*, said P10, *"because... they're still under development."* P11 elaborated:

> *The limitation to protecting [consumer privacy] is the pace of change. I think it outpaces how quickly governments can respond and implement laws and policies... by the time [laws] roll out and by the time new technologies or new areas that affect privacy take place, there's often a lag period before policies are made. (P11)*

This group recognized that the online space is *"the most difficult place to protect Canadians... "If we were to be a hundred percent protected"*, retorted P31, *"[the government] would have to be passing new laws every day."* Consumers' lack of awareness for protecting their privacy is partially due to *"the combination of this old and new technology and people's [lack of] understanding of how it works and what they need to do"* (P27).

Our participants are unsure of how to "fix" privacy concerns under existing and new technology, but instead recommend improved usability and access to privacy resources, tools, and information to keep consumers better informed about their data. Suggestions included displaying information in "more accessible" and "user-friendly" formats. For example, reducing lengthy privacy policies to succinct summaries or short videos. From a utilitarian perspective, some envisioned more accessible ways to find their personal information held by companies in a centralized database where *"I could access which companies have information on me and how long are they able to hold it for"* (P11).

C4. Hackers: "Hackers" from both outside and inside companies (e.g., malicious employees) were seen as a significant threat and limitation in protecting consumer data. Companies *"try the best they can,"* explained P2, *"they try to encrypt it, they try to protect it, but there's always someone that can get their hands on [the data], [and] you can never find out who this person is."*

In the face of a data breach, some believe it is *"not really the company's fault that the leak even happened"* (P2). Hackers *"who want your data are willing to go to extreme lengths to get it... trying to stay one step ahead is difficult if not impossible task for a lot of companies, especially medium to small companies"* (P23). Staying ahead of the hackers is an arms race: *"[companies] got to stay one step ahead of the hackers... [It] requires them hiring people that would be hackers... It's kind of like hackers against hackers trying to stay one step ahead of people trying to steal the data"* (P13).

Part of the problem is the lack of security expertise to safeguard consumer data. For example, P3 said, *"small business owners, they start up a website and they take credit card payments through it, but they don't make sure that their website is secure."* Some believed a lack of security expertise to protect against data breaches is not unique to small companies. P27, a part-time auditor, declared to *"have both identified and read about audit findings that are simply mind-boggling, not for small organizations, but for Fortune 500 organizations... Organizations believe they are secure, but in reality they have huge cracks in their security walls."*

Since most believe it is impossible to completely safeguard against hackers, companies should simply *"do everything that they can... to ensure that it's harder for people to break into their system"* (P9). These participants believe it is in the companies' best interest to safeguard consumer data against hackers to uphold their reputation and *"continue to have a good name"* (P12).

6 A moral code for data privacy

We found that participants do not rely on legal guidelines to determine whether what companies do with their data is appropriate. Instead, they weigh the severity of the violation against their own 'moral code' centred on what they feel is right and wrong. Legally compliant conduct is not necessarily interpreted as ethical and moral, nor as protective of the autonomy and privacy and consumers. P21 clearly describes the boundary between legal and ethical conduct in response to the scenario *S8-openness-of-collection*:

> *Do I think that they acted appropriately under the law? I'm hard-pressed to say it's illegal. I mean, I could be wrong on that one, but I don't think that they're being particularly ethical. You know, the fact that you have to kind of jump through hoops to be able to not provide your credit card information for something that's free, that's a little concerning to me... I don't like the optics of it, but are they being unlawful by asking for your credit card information, even for free stuff? To my knowledge, I don't think it's unlawful. (P21)*

Our participants repeatedly identify this boundary between legal and ethical as the "grey zone" where the companies' actions could be technically legal but unethical. As P6 explained in response to S1-outsourcing-abroad: *"I think that's a bit of a grey zone... I think what they did is a kind of a grey area where they can't really be prosecuted or have anything really done to them. I think they acted accordingly, I want to say, but I really don't approve of it."* This 'grey zone' our participants described is based on their perceptions that laws, by definition, are vague with many loopholes that businesses could take advantage of:

> *Just because the law can be vague. I think that it's written that things need to be transparent, and technically [what they did] does fall under the definition of being transparent. Now whether it's moral is a whole other story. I think they've found a loophole in the wording of the law that makes it advantageous for them. They'll end up with more people on their marketing list if they do it the way they're doing it... (P3)*

.

Pillars	Moral Code	Sub-Codes
I	Trust	Intuition, Reputation, Size, Security Expertise
II	Transparency	Honesty, Purpose, Best Interest
III	Control	Choice, Consent
IV	Access	Access, Usability, Recourse

Table 3: The components of the moral code

This sets the stage for the last step of our Grounded Theory analysis. We propose that participants' understanding and perspective follows a "moral code" for data privacy. We based our model on the identified codes, patterns, and relationships between concepts identified in the analysis. We refined the results into four core values that consumers use to navigate their information disclosure: trust towards the organizations, transparency of the organization, feelings of control over personal information, and access to privacy information. We summarize the components of the moral code in Table 3. Participants' responses to the ten privacy scenarios from Table 2 offer examples of the moral code in practice.

6.1 Pillar I: Trust

Trust towards companies strongly influenced our participants' perceptions of whether the companies' privacy conduct was appropriate, with some participants weighing privacy decisions entirely based on trust towards the business.

Intuition: In judging privacy violations, participants relied primarily on their gut feelings towards a situation. For instance, P5's response to *S1-outsourcing-abroad*,

> *I feel really uncomfortable about that situation because it feels like they're holding your data hostage and switching you from the country with laws that you initially signed up for... to a whole different system that you might not be familiar with. I would assume that they're following the law because at least they're informing you... (P5)*

Participants used words like "red flag", "creepy", "sketchy", "annoyed", "sneaky", "uncomfortable", and "suspicious" to describe questionable conduct. P6 admitted: *"[the situation] just seems kind of sketchy to me. You know, it's not a very academic term... but it kind of rubs me the wrong way."*

Reputation: We avoided naming specific companies in the scenarios, but some participants indicated that their attitude towards a privacy violation would depend on the business. For example, in response to *S1-outsourcing-abroad*, P2 explained that they wouldn't be concerned if the company was Google because *"They're reputable,"* while others expressed distrust if the business was Facebook under the same scenario. Similarly, in *S3-opt-out-consent*, P3's interpretation of whether the business acted appropriately under the law would *"depend on my company." "If it were a reputable company,"* continued P3, *"I wouldn't be concerned... [If it's] a brand new cell phone company, I would be a little bit concerned because they don't have the reputation... to protect my data."*

Some participants defaulted to trusting reputable companies. P31 explained, *"I deal with companies that I believed to be reputable. So I would assume that they're following the rules and the regulations and doing things properly... I assume they're not breaking the law."*

Size: Our participants perceived larger companies to be more trustworthy. P2 explained: *"Bigger companies just have a standard to live by... "* Others shared similar opinions, such as *"a large company... would know better"* (P8). They believed that larger companies have *"a human resource person or someone who's appointed to deal with privacy and legal issues,"* and are, therefore, *"better informed than a small [company], who may not have the staffing to deal with [privacy and legal issues]"* (P8). Smaller companies may be *"not be as compliant... [because they] just don't have the professional expertise to know what the law is exactly"* (P8).

Security expertise: Some participants also believed that small to medium-sized companies lack security expertise for protecting data against hackers. This is because *"even experts have to continually keep up with hackers who are, you know... have a lot of incentive and they may be very well educated and capable people, more so than the actual people who were dealing with the security for the company... only the largest companies with deep pockets could afford to get an adequate level of security"* (P8). These participants shared the view that even though they may not like the idea of sharing certain information with businesses, they felt more at ease with sharing their data with large companies because they perceived them to be better equipped to protect their data. P21 explained in response to scenario *S8-openness-of-collection*:

> *The fact that you said that it's a fairly well-known company asking for the information, I feel fairly safe that they're going to protect my personal information. I mean, ultimately, any company is going to be at risk of being hacked or having their information taken from them. But I usually feel a lot safer when it's like a big company versus it being, you know, someone smaller like fly-by-night. (P21)*

Participants thus believe that privacy protections and standards vary significantly across different organizations, and they generally placed greater trust in larger companies.

6.2 Pillar II: Transparency

Whether a business is transparent and forthcoming about its conduct influenced our participants' assessment of the severity of privacy violations and their acceptance of an organization's privacy practices.

Honesty: Being honest and forthcoming were identified as essential values. Obfuscating privacy-compromising practices is viewed as dishonest and unethical. In P12's words, *"I feel like it's dishonest. I don't think it's the most ethical thing. I think companies are always out to sort of serve their own*

interests. And if it's not in their interests for you to be aware of all that information, they're not going to make it always easy for you to find." P23 described the concern further in *S8-openness-of-collection*, *"[the company is] hiding important information in spots where, you know, vulnerable, uneducated, unknowing people would never [look], would never see... I think that's sleazy... why hide that information?"*

Many of our participants were willing to forgive certain types of misconduct if the organization is honest about it. For example, in the event of a data breach in *S7-safeguarding-data*, many participants believed that the recovery effort is redeemable because the business did not try to cover up the breach. P4 explained, *"security breaches happen. So I wouldn't fault them for the actual security breach. If afterwards, they do everything to try and deal with the breach appropriately, then that's fine..."*

Purpose: Participants showed greater comfort and acceptance towards data collection if they understood and agreed with its purpose. For example, in *S2-GPS-tracking*, most thought it reasonable to track company vehicles because they are the company's property and not an employee's private space. Hence it was not considered an intrusion of privacy.

From a legal perspective, the business in *S4-over-collection "said what information they needed, explained what they're going to do with it, and [said] they're not going to keep it past that time... which keeps all within the guideline"* (P32), but our participants felt uneasy about *"whether [the company] actually needed that information in the first place"* (P32). *"A grocery store doesn't need my driver's license number or utility bill for... confirming who you say you are"* (P24).

Best interest: If a business has acted with the consumers' best interest in mind, our participants view the actions as ethical. In the case of being denied access to online accounts, P21 said, *"I've had that issue with an e-mail address being hacked previously... and jumped through a whole lot of hoops to get [my] account back... they ask for a lot of information that maybe shouldn't be necessary. But ultimately, I think they're trying to protect the consumer. They're trying to ensure that you are actually you."* In a similar situation in *S9-accessing-password* where the business denied the individual's access, our participants rationalized that the business acted responsibly from an ethical point of view. *"It sucks that I have to jump through all these hoops to get my answer"*, responded P22, *"but it sounds like they're doing a better job of respecting and looking after my data."* P31 agreed, *"seeing that I tried to get information from them and they said 'no'... I'd probably actually feel better about it. So I'd change my password and wouldn't feel concerned."*

6.3 Pillar III: Control

Our participants felt that they lacked control over their personal information once a company collects it. For example, P28 said in response to *S4-over-collection*, *"I just have this feeling that once you send this information, you really have no idea what they're doing with it. Like they're saying they're going to do that. But you have no idea what actually happens to it after."* P22 agrees, saying that *"Once you enter [your information] it goes to this kind of black hole of not knowing... What do they do with it? You're kind of at their leisure, at their discretion."*

Choice: When asked about why they would give their personal information when feeling uneasy doing so, our participants identified a lack choice for the services and products they need as one of the main reasons. P1 explained, *"I don't even know if I'm on any Canadian servers because most of the stuff we use is in the US and beyond... I don't know the alternatives. I mean, if I went looking for Canadian alternatives to the services I use, I suspect I wouldn't find that many [laughs]."* P1 continued, *"you gotta pick the Apple or Microsoft or Google these days because it's pretty much the 3 things that make devices and software to put on them,"* P7 complained, *"You sort of need to sign away your rights to be able to do things."* Our participants felt cornered when organizations try to provide the perception of choice and control over personal information. In *S1-outsourcing-abroad*, P5 felt *"they're holding your data hostage"* if the customer does not agree to the terms. These participants felt uneasy giving away their information but believed they had no other options. In the words of P11, *"I kind of went into a spot where I didn't necessarily have an alternative option, so I complied, but I just kind of didn't like it."*

Consent: All participants agreed that obtaining consumers' consent before data collection is the basis of lawful conduct. Several participants held the view that consent should always be explicit. "Opt-out" consent was viewed as being unethical practice. P23 explained in response to *S3-opt-out-consent*:

> *I don't know if the law is an "opt-out" or an "opt-in" type of law, but... but I don't think they acted ethically... It shouldn't be like, hey, if you don't want this, then you have to do, you know, jump through the hoops in order to make sure that you don't want this. It should be. Hey, if you want to be included, give us a call... and [opt-out] is not the way that consent works, nor should it work that way. (P23)*

Similarly, in *S5-amending-consent*, P19 felt the business should *''get the confirmation from customers that they feel comfortable with [the changes in the terms and conditions] rather than letting them know that, you know, we're [already] doing that."* Consent should, therefore, be *"brought about by the individual, and the individual should be the one to make the decision–full stop"* (P23).

6.4 Pillar IV: Access

Our participants wanted *"a better sense of accessibility of [their] data"* (P5), including access to how companies manage their personal information, more usable privacy informa-

tion, and clear recourse for addressing privacy concerns.

Access: Our participants identified that they lack access to details about how companies handle their personal information, what information companies have about them, where their data is stored, how long the data is kept, and when it gets destroyed. P8 recounted their experience requesting access to personal data:

> *I have in a couple of instances tried to contact companies about what information they have about me, and had some positive replies in terms of they've given me the information, sent me the information, or said that they would delete the information. Although I can't guarantee that it's gone. At least, they said they would. I have also received no response from some places, in which case I assume that they're probably not deleting it. And then also there's the case of companies that go under and you can't... I tried to contact [a company] that I knew had quite a bit of information about me and they're gone, but it doesn't mean their databases are gone. If a company goes bankrupt or something. I think in many times in a lot of those things are just ignored. So where are they kept? Where's the servers, and what happened to them? Did anybody ever delete it properly? Did the hard drive just get thrown into the garbage somewhere? (P8)*

In response to the scenario *S7-safeguarding-data*, P9 raised the concern that *"we have no idea what's happening with our information. The only time that we ever find out... that something is wrong is when there is a big announcement that the information was breached and this many customers were affected... But apart from that... I don't feel like I know anything"* (P9). Denying consumers access to their personal information could erode trust. In *S10-challenging-exception*, P13 believe the doctor *"acted appropriately under the law, but don't think that it's right that there are notes about you that you're not allowed to see."*

Usability: Unsurprising, participants' lack of awareness is partially due to not reading terms and conditions before signing up for a service or product because they are "absurdly long". Even those who are privacy-conscious find it challenging to understand privacy policies. After experiencing a data breach, P23 said, *"I started being more aware of privacy and who I give my information to and even going as far as looking at companies policies as to their storage of user data. And a lot of it's, you know, I would say, verbose. Like it's not really clear on what they're doing with their with your data or information, you're sort of just asked to trust them unilaterally."* When responding to scenarios like *S3-opt-out-consent* and *S8-openness-of-collection*, Our participants are conscious that companies recognize that most people don't read policies and take advantage of the "loopholes" in getting users to agree to their terms and services. *"I feel like it's dishonest,"* said P12.

Recourse: While our participants realize they have legal privacy rights and that businesses are under certain obligations to protect consumer privacy, barriers exist that prevented participants from identifying and challenging a business who infringes on their privacy. For those who had raised a concern, many did not have a satisfying resolution. Several of the companies our participants contacted did not follow up to confirm whether the concern was addressed. P32 said, *"I emailed the company, and they called me, and I actually spoke to... their supervisor... they assured me they would sit down and look at their process and see if there were anything they could do... at least... they said they would (laughs), but I don't know what happened after that."* When a company doesn't follow-up, people tend to give up and *"just let it go"* (P19). Understandably, some of our participants had *"a lack of faith in the system that something is going to be done"* (P23) if they raised a concern, and they *"don't trust companies to be as accountable as they should be."* (P26) Aside from not knowing whom to report concerns to, P1 elaborated, *"as far as if they would actually do anything... like what do you do? Go to the police and tell them Facebook's not doing what you asked them to? ... There's nothing really clear beyond just going to the company and hoping they actually listen to you, which they usually don't."*

7 Discussion and future work

A commonality between many existing privacy theories is that individuals' perceptions of privacy depend on situational circumstances. Privacy regulations like PIPEDA define privacy through regulations for controlling the flow of personal information about individuals. We suggest that there exists a misalignment between privacy regulations based on FIPs and privacy theories like contextual integrity [23], organizational trust [22], Solove's Taxonomy of Privacy [35], and our concepts of Moral Codes. These works show that preserving privacy is not only a matter of controlling the flow of personal information, but also how privacy practices and norms meet individual and societal values. Our work contributes to identifying specific moral values that individuals abide by in making privacy decisions.

The Government of Canada has recently suggested changes to PIPEDA in conjunction with the Digital Charter that specifically mandates "the ethical use of data to create value, promote openness and improve the lives of people—at home and around the world" as one of the guiding principles [15]. As indicated by our results, PIPEDA's FIPs focusing on the basic technical and legal responsibilities of organizations are insufficient to address the ethical and ecological concerns that emerge and ascend to the top of minds during consumers' privacy decision-making. Our participants' Moral Codes suggest new rights and expectations for privacy, including increased access, meaningful choices, clearer information, the ability to move or remove information, and real accountability through stronger enforcement. Based on our study results, we propose the following recommendations.

Consent Model: Most participant concerns center on

PIPEDA's current model of "implied" consent that allows businesses to claim they have an individual's consent to use their information in a certain way without asking for it. For example, the cellular company in the compliant scenario *S3-opt-out-consent* could argue it has "implied" consent because they are using existing customers' information and, by signing up for the service, customers must have implied consent to receive marketing material. Our participants deemed this approach within the boundary of the law but highly unethical. This observation suggests that mismatches between corporate privacy practices and individuals' personal values were likely to be viewed as unethical. Many participants referenced the GDPR as a model they would like to see incorporated into Canadian law. An organization under GDPR must have "legitimate interests" to use personal information, such as fraud prevention. Our results suggest that this model is in closer alignment with the Moral Codes that consumers abide by, such as *Purpose* and *Best interest*. Therefore, we recommend adopting a consent model similar to GDPR's "legitimate interest" model to replace the "implied consent" model in PIPEDA.

Control and Access: Descriptions of the FIPs appeared to satisfy participants' moral expectations superficially, but in practice, they were disappointed with their weak enforcement and vague applicability to real-life privacy situations, leaving individuals powerless to control and access their personal information. For example, many participants felt "trapped" and like they had no choice but to agree to *S1-outsourcing-abroad* for fear of losing their data. PIPEDA provides "right of access" and limited "right to deletion" of inaccurate or outdated personal information. Our participants also desired stronger rights to deletion and the "right to data transfer", where they could request their personal information in an accessible and portable format to transmit it to a different organization. However, usability testing needs to be conducted on which data formats (e.g., CSV, JSON, XML) are more usable and accessible to end-users, possibly developing new human-readable formats. Other usability issues that create barriers for control and access identified by our participants, such as the presentation of privacy policies and privacy settings, could be addressed by standardizing certain key interface elements. For example, the State of California Department of Justice has released a standard "Privacy Options Opt-Out Icon" to direct users to opt-out [36].

Assessment Tools: Our results suggest participants were ill-equipped to identify privacy violations and hold businesses accountable using legal frameworks like the PIPEDA. Instead, they relied on their own moral assessment of businesses' privacy conduct based on trust, transparency, control, and access. Therefore, we suggest using the Moral Codes as a framework to develop tools that help organizations align their practices and policies with consumer expectations. For example, the Information Commissioner's Office (ICO) in the UK has developed a three-part test [40] with an ethics component to help businesses determine whether they have a legitimate interest in processing consumers' personal data.

We further propose that an independent entity such as the Better Business Bureau [9] could conduct an assessment and provide ratings for organizations on the basis of their privacy practices. Our Moral Codes could be used as one of the criteria guiding this type of assessment. This would enable customers to seek out organizations that meet their privacy expectations and may serve as incentives for organizations to improve their practices.

Despite this potential incentive, a key problem lies with how to convince corporate organizations to take these steps. Competing corporate priorities mean that there is little incentive for them to prioritize "moral" or privacy-preserving designs, and in many cases, there are significant economic and competitive disincentives. Our view is that this issue requires increased governmental regulation and oversight, and only once this is in place will there be sufficient interest in making practical changes. However, studies such as this one help increase awareness among stakeholders, and provide supporting evidence to those in positions to push for change.

Limitations: We chose to present participants with scenarios and information about the privacy principles, which may have primed them and increased their privacy concern. This was a considered methodological choice because we wanted participants to engage with the principles and provide their perspectives, but we also knew from background research that people were likely unfamiliar with the principles. Our survey opened with demographic questions Westin's Privacy Segmentation Index to compare the overall privacy attitudes from our sample to previous studies. As indicated by some studies (e.g., [43]), the Westin categories may not accurately infer behavioural intent and responding to demographic questions first could increase the stereotype threat [38]. Our work is focused on users and regulations from Canada; while we broadly think that our findings would generalize, at least to other Western countries, further work is needed to explore the unique attributes present in other parts of the world.

8 Conclusion

Making online privacy decisions is increasingly difficult due to the complexity of information technologies and the variety of activities that consumers engage with online across multiple platforms and devices [3]. Our research adds to the body of literature in understanding individual's privacy preferences and behaviours. Beyond the traditional economic view of individuals engaging in privacy benefit trade-offs, and heuristics and biases that influences behaviour, we suggest that understanding users' privacy ethics could offer rich insights into how they engage in online privacy decision-making.

Acknowledgments

Sonia Chiasson acknowledges funding from NSERC for her Canada Research Chair and Discovery Grants. The authors thank Elisa Kazan for help in data collection and Cassie Cassell for help in qualitative analysis.

References

[1] Eathar Abdul-Ghani. Consumers' online institutional privacy literacy. In *Advances in Digital Marketing and eCommerce*, pages 40–46. Springer, 2020.

[2] Mark S Ackerman, Lorrie Faith Cranor, and Joseph Reagle. Privacy in e-commerce: Examining user scenarios and privacy preferences. In *ACM conference on Electronic commerce*, pages 1–8, 1999.

[3] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, et al. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys (CSUR)*, 50(3):1–41, 2017.

[4] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.

[5] Alessandro Acquisti and Jens Grossklags. Privacy and rationality in individual decision making. *IEEE security & privacy*, 3(1):26–33, 2005.

[6] Irwin Altman. *The environment and social behavior: privacy, personal space, territory, and crowding.* Brooks Cole Publishing, 1975.

[7] Oshrat Ayalon and Eran Toch. Evaluating users' perceptions about a system's privacy: Differentiating social and institutional aspects. In *Symposium on Usable Privacy and Security (SOUPS)*, pages 41–59, 2019.

[8] Paula J Bruening and Mary J Culnan. Through a glass darkly: From privacy notices to effective transparency. *NCJL & Tech.*, 17:515, 2015.

[9] Butter Business Bureau. BBB start with trust, 2021.

[10] Sunny Consolvo, Ian E Smith, Tara Matthews, Anthony LaMarca, Jason Tabert, and Pauline Powledge. Location disclosure to social relations: why, when, & what people want to share. In *SIGCHI conference on Human factors in computing systems*, pages 81–90, 2005.

[11] Juliet Corbin and Anselm Strauss. *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications, 2014.

[12] Curt J Dommeyer and Barbara L Gross. What consumers know and what they do: An investigation of consumer knowledge, awareness, and use of privacy protection strategies. *Journal of Interactive Marketing*, 17(2):34–51, 2003.

[13] Janna Lynn Dupree, Richard Devries, Daniel M Berry, and Edward Lank. Privacy personas: Clustering users via attitudes and behaviors toward security practices. In *SIGCHI Conference on Human Factors in Computing Systems*, pages 5228–5239, 2016.

[14] Nina Gerber, Paul Gerber, and Melanie Volkamer. Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. *Computers & Security*, 77:226–261, 2018.

[15] Government of Canada. Canada's digital charter: Trust in a digital world, 2021.

[16] David R Hannah and Brenda A Lautsch. Counting in qualitative research: Why to conduct it, when to avoid it, and when to closet it. *Journal of Management Inquiry*, 20(1):14–22, 2011.

[17] Lesley Jacobs, Barbara Crow, and Kim Sawchuk. Privacy rights mobilization among marginal groups in canada: Fulfilling the mandate of PIPEDA. Technical report, York Centre for Public Policy & Law, York University, 2011. https://ycppl.info.yorku.ca/files/2013/05/Privacy-Rights-PIPEDA-paper.pdf.

[18] Leslie K John, Alessandro Acquisti, and George Loewenstein. Strangers on a plane: Context-dependent willingness to divulge sensitive information. *Journal of consumer research*, 37(5):858–873, 2011.

[19] Spyros Kokolakis. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & Security*, 64:122–134, 2017.

[20] Klaus Krippendorff. Reliability in content analysis. *Human communication research*, 30(3):411–433, 2004.

[21] Klaus Krippendorff. Computing krippendorff's alpha-reliability. *Departmental Papers (ASC)*, 43, 2011.

[22] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of Management Review*, 20(3):709–734, 1995.

[23] Helen Nissenbaum. *Privacy in context*. Stanford University Press, 2009.

[24] Office of the Privacy Commissioner of Canada. PIPEDA interpretation bulletins, 2020.

[25] Leysia Palen and Paul Dourish. Unpacking "privacy" for a networked world. In *SIGCHI conference on Human factors in computing systems*, pages 129–136, 2003.

[26] Robert W Palmatier and Kelly D Martin. *The intelligent marketer's guide to data privacy: the impact of big data on customer trust*. Springer, 2019.

[27] Yong Jin Park. Digital literacy and privacy behavior online. *Communication Research*, 40(2):215–236, 2013.

[28] Pew Research Center. The state of privacy in post-snowden america, 2016.

[29] Ashwini Rao, Florian Schaub, Norman Sadeh, Alessandro Acquisti, and Ruogu Kang. Expecting the unexpected: Understanding mismatched privacy expectations online. In *Symposium on Usable Privacy and Security (SOUPS)*, pages 77–96, 2016.

[30] Kate Raynes-Goldie. Aliases, creeping, and wall cleaning: Understanding privacy in the age of facebook. *First Monday*, 15(1):1–14, 2010.

[31] Marshall David Rice and Ekaterina Bogdanov. Privacy in doubt: An empirical investigation of canadians' knowledge of corporate data collection and usage practices. *Canadian Journal of Administrative Sciences/Revue Canadienne des Sciences de l'Administration*, 36(2):163–176, 2019.

[32] Katharine Sarikakis and Lisa Winter. Social media users' legal consciousness about privacy. *Social Media & Society*, 3(1):1–14, 2017.

[33] Kim Bartel Sheehan. Toward a typology of internet users and online privacy concerns. *The Information Society*, 18(1):21–32, 2002.

[34] H Jeff Smith, Tamara Dinev, and Heng Xu. Information privacy research: an interdisciplinary review. *MIS quarterly*, pages 989–1015, 2011.

[35] Daniel J Solove. A taxonomy of privacy. *University of Pennyslvania Law Review*, 154:477, 2005.

[36] State of California Department of Justice. California consumer privacy act (CCPA) opt-out icon, 2021.

[37] Statistics Canada. 2016 census profile, 2017.

[38] Claude M Steele and Joshua Aronson. Stereotype threat and the intellectual test performance of african americans. *Journal of personality and social psychology*, 69(5):797, 1995.

[39] S Shyam Sundar, Jinyoung Kim, Mary Beth Rosson, and Maria D Molina. Online privacy heuristics that predict information disclosure. In *SIGCHI conference on Human factors in computing systems*, pages 1–12, 2020.

[40] TermsFeed. 3 part test for legitimate interests under the GDPR, 2021.

[41] Edward Shih-Tse Wang. Role of privacy legislations and online business brand image in consumer perceptions of online privacy risk. *Journal of theoretical and applied electronic commerce research*, 14(2):0–0, 2019.

[42] Alan F Westin. Privacy and freedom. *Washington and Lee Law Review*, 25(1):166, 1968.

[43] Allison Woodruff, Vasyl Pihur, Sunny Consolvo, Laura Brandimarte, and Alessandro Acquisti. Would a privacy fundamentalist sell their DNA for 1000... if nothing bad happened as a result? the westin categories, behavioral intentions, and consequences. In *Symposium On Usable Privacy and Security (SOUPS)*, pages 1–18, 2014.

[44] Heng Xu, Tamara Dinev, Jeff Smith, and Paul Hart. Information privacy concerns: Linking individual perceptions with institutional privacy assurances. *Journal of the Association for Information Systems*, 12(12):1, 2011.

APPENDIX

A. Online Survey

The survey choices are formatted differently than what appeared in the Qualtrics survey seen by participants to conserve space.

A1 Demographic Questions

Q1. Which province or territory are you currently living in? (Choices: The thirteen provinces and territories)

(The following questions all include a "Prefer not to answer" choice.)

Q2. Which gender do you identify as? (Choices: Male, Female, Non-binary, Other)

Q3. What age group do you belong to? (Choices: 19 years and under, 20 years to 79 years in five-year intervals, 80 years and above)

Q4. What is your highest level of education? (Choices: Less than a high school degree, High school degree or equivalent, College degree, Bachelor's degree, Master's degree, Doctoral degree, Other professional degree)

Q5. What is the total income of your household per year? (Choices: Less than $15,000, $15,000 to $99,999 in $4,999 intervals, $100,000 to $149,000, $150,000 to $199,999, $200,000 and above)

A2 Westin privacy index Questions

Q6. Participants responded to the following questions with a four-point scale ranging from "Strongly Agree" to "Strongly Disagree"

1. Consumers have lost all control over how personal information is collected and used by businesses.
2. Most businesses handle the personal information they collect about consumers in a proper and confidential way.
3. Existing laws and business practices provide a reasonable level of protection for consumer privacy today.

A3 Survey questions

Q7. Which, if any, of the following types of Internet-connected device(s) do you have in your household?

1. Mobile phones
2. Tablets
3. Desktop or laptop computers
4. Smart appliances (e.g., gas/electric meters, refrigerators, thermostats or robotic floor cleaners)
5. Smart media devices (e.g., printers, speakers, TVs)
6. Wearables (e.g., smartwatches, Fitbit)
7. Medical health monitors (e.g., Smart continuous glucose monitoring (CGM) and insulin pens, smart inhalers, smart heart monitors)
8. Home assistants (e.g., Amazon Alexa or Google Assistant)
9. Gaming consoles connected to the Internet (e.g., Xbox, PlayStation 4 or Nintendo Wii U)
10. Home security systems connected to the Internet (e.g., SimplySafe)
11. Toys, baby monitors or GPS child trackers connected to the Internet (e.g., Hello Barbie, Furby Connect, Phillips Avent, Amber Alert)
12. Car with smart system (e.g., Audi Connect, Lexus Enform, Ford SYNC3)

Q8. How would you rate your knowledge of your privacy rights? (Choices were a five-point scale ranging from "Very good" to "Very poor")

Q9. How would you rate your knowledge of how to protect your privacy rights? (Choices were a five-point scale ranging from "Very good" to "Very poor")

Q10. In general, how concerned are you about your personal information held by businesses? (Choices were a five-point scale ranging from "Very concerned" to "Not at all concerned")

Q11. Participants responded to the following questions with a four-point scale ranging from "I am aware for *all* of the services and products that I use" to "I am aware for *none* of the services and products that I use"

1. What personal information is collected and its sensitivity
2. Why my personal information is collected
3. How my personal information collected
4. What my personal information is used for
5. Where my personal information is physically kept
6. How my personal information is protected and secured
7. Who has access to or uses my personal information
8. Who my personal information is shared with
9. after it is no longer needed

Q12. In general, how would you rate your knowledge of how these technologies affect your privacy? (Choices were a five-point scale ranging from "Very good" to "Very poor" with a "Don't know" option)

Q13. Participants rated their knowledge of how to protect their personal information on the following Internet-connected devices using a five-point scale ranging from "Very good" to "Very poor" with a "Don't know" option.

1. Mobile phones
2. Tablets
3. Desktop or laptop computers
4. Smart appliances (e.g., gas/electric meters, refrigerators, thermostats or robotic floor cleaners)
5. Smart media devices (e.g., printers, speakers, TVs)
6. Wearables (e.g., smartwatches, Fitbit)
7. Medical health monitors (e.g., Smart continuous glucose monitoring (CGM) and insulin pens, smart inhalers, smart heart monitors)
8. Home assistants (e.g., Amazon Alexa or Google Assistant)
9. Gaming consoles connected to the Internet (e.g., Xbox, PlayStation 4 or Nintendo Wii U)
10. Home security systems connected to the Internet (e.g., SimplySafe)
11. Toys, baby monitors or GPS child trackers connected to the Internet (e.g., Hello Barbie, Furby Connect, Phillips Avent, Amber Alert)
12. Car with smart system (e.g., Audi Connect, Lexus Enform, Ford SYNC3)

Q14. For each of the statements, how would you rate your knowledge regarding your privacy?
(Choices were a five-point scale ranging from "Very good" to "Very poor" with a "Don't know" option)

1. The basics of Canada's federal privacy laws
2. How the Federal Government handles my personal information
3. A business's obligations concerning my privacy and personal information
4. How to raise a privacy concern with businesses that handles my personal information
5. How to file a privacy complaint with a business to the Office of the Privacy Commissioner of Canada (OPC)

Q15. To what extent do you agree or disagree with the following statements for protecting your privacy? (Choices were a five-point scale ranging from "Always" to "Never")

1. I think about why my personal information is needed, who will use it, and how it would be used before providing it online or in person.
2. I read the privacy policies of the websites and apps I use
3. I raise my concerns with the business if I am worried about the way my personal information is being handled.
4. I refuse to provide optional personal information when a business asks me for it. (e.g., when a business asks you to provide an optional secondary phone number).
5. I remove my personal information when I no longer need the services that I signed up for (e.g., removing yourself from mailing lists).
6. I avoid sharing my Social Insurance Number (SIN) with businesses or individuals (e.g., landlords).

7. I ensure my computer, smartphone and other mobile devices are password protected.
8. On my devices, I download from reputable sources.
9. On my devices, I install the latest software updates.
10. On my devices, I encrypt sensitive data.
11. On my devices, I disable Wi-Fi and Bluetooth if I'm not using it.
12. On my devices, I disable Wi-Fi and Bluetooth when passing through public spaces with open wireless networks.
13. I create passwords that are sufficiently complex using character combinations that are only meaningful to me.
14. I use different passwords for different websites, accounts and devices.

Q16. To what extent do you agree or disagree with the following statements for protecting your privacy? (Choices were a five-point scale ranging from "Strongly agree" to "Strongly disagree")

1. I regularly review and adjust the privacy settings on my devices to limit the sharing of my personal information with businesses.
2. In general, I believe Canadian privacy laws effectively protect my privacy.

A4.1 Privacy Scenarios

(Each participant was randomly assigned to five out of ten scenarios. The order was randomized. One scenario was displayed per page).

Q17. There are privacy principles for businesses to comply with the law regarding how they collect, use and disclose individuals' personal information. The next 5 questions will include various scenarios about the privacy practises of businesses. Imagine yourself in each of the following scenarios and indicate to what extent do you agree or disagree with each statement. (Choices were a five-point scale ranging from "Strongly Agree" to "Don't know" with a "Don't know" option)

1. I think scenarios like this are likely to happen.
2. I would be concerned about my privacy in this scenario.
3. I think the business acted appropriately in a lawful manner based on the situation described.
4. Which of the privacy principles do you think apply in this situation? (The principles were displayed as a checklist)
    a. **Accountability:** A business is responsible for personal information under its control. It must appoint someone to be accountable for its compliance with these privacy principles.
    b. **Identifying Purposes:** The purposes for which the personal information is being collected must be identified by the business before or at the time of collection.
    c. **Consent:** The knowledge and consent of the individual are required for the collection, use, or disclosure of personal information, except where inappropriate.
    d. **Limiting Collection:** The collection of personal information must be limited to that which is needed for the purposes identified by the business. Information must be collected by fair and lawful means.
    e. **Limiting Use, Disclosure, and Retention:** Unless the individual consents otherwise or it is required by law, personal information can only be used or disclosed for the purposes for which it was collected. Personal information must only be kept as long as required to serve those purposes.
    f. **Accuracy:** Personal information must be as accurate, complete, and up-to-date as possible in order to properly satisfy the purposes for which it is to be used.
    g. **Safeguards**: Personal information must be protected by appropriate security relative to the sensitivity of the information.
    h. **Openness:** A business must make detailed information about its policies and practices relating to the management of personal information publicly and readily available.
    i. **Individual Access:** Upon request, an individual must be informed of the existence, use, and disclosure of their personal information and be given access to that information. An individual shall be able to challenge the accuracy and completeness of the information and have it amended as appropriate.
    j. **Challenging Compliance:** An individual shall be able to challenge a business's compliance with the above principles.

A4.2 Scenario description

All scenarios are based on real reported findings from OPC's Interpretation Bulletins, linked at the end of each scenario. We shortened the case summaries and retained only the essential information in a standard format to maintain consistency and improve readability.

1. You received an email from your Canadian email provider notifying you that your email services would be operated by a business based in the U.S from now on. The email provider is informing you that your data will be used and stored in the U.S., which is subject to the laws of that country. The email states that upon logging into your new account, you will be asked to accept or decline the new services. If you decline, your email account and all its contents will be permanently deleted. *Based on PIPEDA Case Summary #2008-394*
2. You are an employee of a telecommunications company that does installation and repairs. Your employer notifies you that they are installing Global Positioning Systems (GPS) on all work vehicles to manage workforce productivity, ensure safety and development, and protect and manage assets. The GPS data will be used to locate, dispatch, and route employees to job sites. Your employer will be able to view and track the location of your vehicle in real-time and to produce reports using historical data. *Based on PIPEDA Case Summary #2006-351*
3. You receive a privacy brochure as an insert in your monthly cellular telephone bill. The brochure outlines the business's intended practices regarding the collection, use, and disclosure of customers' personal information for secondary purposes of marketing, and lists all parties concerned. The brochure also indicates that you could have your name removed from marketing lists by calling a toll-free number, sending an email, or using the business's website. If you do not notify the business of your intention to withdraw, it will assume your consent to the continued collection, use, and disclosure of personal information for the identified purposes. *Based on PIPEDA Case Summary #2003-207*
4. A business is offering you a free $10 Grocery Card to purchase items sold in their grocery stores. While registering for the Grocery Card, the business notifies you that to confirm that they are issuing a $10 Grocery Card to a single eligible person, you are required to provide a scanned copy or photo of either: (i) a current utility bill or (ii) a valid driver's licence to finish processing your registration. You are told that the information will not be used for any purpose other than to verify your eligibility and will be destroyed as soon as the verification is complete. *Based on PIPEDA Case Summary #2019-003*
5. You receive a notice from your bank that it is amending its personal information consent clause for its credit and deposit agreements. The notice explains that the amendment is to notify customers that the bank intends to use their personal information for the secondary purpose of marketing new products and services. It also includes a note about who would have access to customers' personal information. The form indicates that customers can withdraw consent by contacting the bank, although it warned that doing so might restrict the bank's ability to effectively provide products and services. *Based on PIPEDA Case Summary #2003-192*
6. You found out that a fraudster had opened a store credit card account with your bank using your personal information. The bank stated that the applicant presented false identification and completed the application form. The form included name, date of birth and SIN, which appear to have been yours. In addition, the address provided was very similar to your address. The bank's credit representative was suspicious and alerted its security department about the account. The security department of the bank made attempts to contact you by telephone using the information on file, but the attempts were unsuccessful. The fraudster used your information to obtain a store credit card and bought $9,000 worth of goods. You contacted your bank to initiate an investigation and to flag the charges as fraudulent. Your bank assumed the financial loss for the account balance. *Based on PIPEDA Case Summary #2007-381*
7. You receive an email notifying you that the server of a web-enabled toy manufacturer, in which you are a customer, was hacked. As a result, there was unauthorized access to account-related information, potentially including your and your children's personal information. The toy manufacturer undertook steps to contain the breach, mitigate the risks to individuals whose information had been compromised, and improve safeguards to minimize the risk of a future breach. *Based on PIPEDA Case Summary #2018-001*
8. You download a free app on your mobile device from a well-known technology company. You are asked to create a User ID for accessing online services before downloading the free application. The registration process includes entering your credit card information. To provide customers with instructions about how to download free applications without having to provide their payment information, the business posted the

information in the website's support section. The information could also be found by using the search term "credit card" in its website's search engine. *Based on PIPEDA Case Summary #2014-007*

9. You attempt to log on to your email account, but your password does not work and you have to reset it. This is the second time it has happened in less than a month and you are suspicious that someone is changing the password to gain access to your account. You contact the business by email, informing them of the problem and requesting access to the date, time, and IP address of the computer being used to change the password. The business replied saying that it cannot grant you access to password information because it is typically law enforcement officials or lawyers who request this information and not clients. The business informed you that if you want information regarding password changes, you would need to provide a subpoena or court order. *Based on PIPEDA Case Summary #2005-315*
10. You contact your doctor asking for a copy of a report that your doctor sent to your insurance company after a medical examination and the written notes that he took during the examination. Your doctor provided you with a copy of the report but refused to provide his notes, indicating that in his view, they did not form part of your medical record, and were therefore not your personal information. The doctor stated he would rely on two exceptions under the law to refuse access: 1) a business is not required to give access to personal information only if the information is protected by solicitor-client privilege; and 2), a business may not give access only if the information was generated in the course of a formal dispute resolution process. *Based on PIPEDA Case Summary #2005-306*

Q18. We are interviewing people about their privacy awareness and experiences. Selected participants can expect the interview to take one hour to complete via a video chat platform (e.g., Skype), and be compensated for their time. If you agree to be contacted about the interview, you will be asked to provide your Prolific ID for sending you study information. Your decision will not impact your payment for the current survey. (Choices: Yes, please email me more information about the follow-up interview, No, I do not wish to be contacted.)

B. Interview

B1 General questions

Q1. What is your definition of "personal information"?
Q2. How do Canadian privacy laws protect the rights and privacy of consumers regarding the collection, usage, and disclosure of their personal information by companies?
Q3. In general, do companies provide reasonable protection for consumers' privacy? Why or why not?
Q4. In general, do existing laws provide reasonable protection for consumers' privacy? Why or why not? Are there any extra protections that you think should exist?
Q5. How did you learn about Canadian privacy protections? Have you ever gone looking for more information about privacy protections? If yes, why did you decide to do this? Did you find what you needed?
Q6. Whose responsibility is it to report privacy concerns/complaints against a company? Who should it be reported to?
Q7. Have you ever had a privacy concern or complaint against a company? If so, what happened? What did you do? What would you do if you had a concern/complaint tomorrow?
Q8. What are the biggest challenges with protecting consumers' privacy?

B2 Privacy Scenarios

The participants were read the same scenarios they responded to in the survey. See Section A.4.2 Scenarios for the description.

- Do you think the company acted appropriately under the law based on the situation described? Why or why not?
- Would you be concerned about your privacy in this scenario? Why or why not?

The participants also answered the following questions corresponding to the scenarios.

- **Scenario 1**. Do you have an example of a time when a company stored your data outside of Canada (e.g., in the US or another country)?
    a. Were you concerned? Why or why not?

- **Scenario 2**. Can you think of a time when a company did not provide a clear explanation about why they were collecting your personal information?
    a. Can you describe what happened?
    b. Did you provide the information anyway? Why or why not?
- **Scenario 3.** Can you think of a time when you felt concerned about the way that a company is obtaining your consent for the collection of your personal information?
    a. Can you describe what happened?
- **Scenario 4.** Can you think of a time when you felt that a company collected more information about you than it was necessary?
    a. Can you describe what happened?
    b. Did you provide the information anyway? Why or why not?
- **Scenario 5.** Can you think of a time when you felt concerned about a company changing its privacy policies to something different than what you initially consented to?
    a. Can you describe what happened?
- **Scenario 6.** Can you think of a time when a company used inaccurate or outdated information about you?
    a. Can you describe what happened?
    b. Were there consequences?
    c. What did you do to improve the situation?
- **Scenario 7**. Can you think of a time when your personal information held by a company was potentially compromised due to a security breach?
    a. Can you describe what happened?
    b. Were there consequences?
    c. What did you do to improve the situation?
- **Scenario 8.** Can you think of a time when you had difficulties finding certain information about a company's privacy practices relating to your personal information?
    a. Can you describe what happened?
- **Scenario 9.** Can you think of a time when you had difficulties accessing your personal information held by a company?
    a. Can you describe what happened?
- **Scenario 10.** Can you think of a time when you raised a privacy concern with a company?
    a. Can you describe what happened?
    b. Did the company address your privacy concern?

C. Supplementary Results

We first identified five salient descriptions of personal information:

1. **Something that I am:** A group of participants described their biological, intellectual, and cultural makeup as their personal information. This included demographic, health, and medical information. Some stated personal beliefs and interests (e.g., political/religious beliefs, hobbies). Few mentioned biometric information.
2. **Something that I use:** Others believed that personal information is extracted from documents issued to a person. It included government-issued ID, contact information, and financial information (e.g., credit card). A person's name, username, and passwords also fall into this category. Participants believe they should carefully protect *this* information against identify theft and fraud.
3. **Something that I have done:** Some described information gathered through online behavioural tracking methods (e.g., browsing history, location data) as their personal information. Participants with this model were aware that organizations use this information to create tailored content like targeted ads.
4. **My "private" information:** Some equated "personal to "private" and included any information that is not disclosed publicly by choice. It is described as "anything pertaining to myself that's not obvious or publicly available" (P5); "things that normal people can't just look up [on Google]" (P6); "anything that happens… in my house" (P21), and "something you wouldn't know unless you were me or a close family member" (P23).
5. **Like a montage:** A few participants believed that seemingly insignificant details about a person could become personal information when pieced together. For example, "a male in a particular setting who makes a certain amount of money\dots and those individual pieces may not be\dots strictly personal information, but all placed together, they become identifiable" (P7).

Pursuing Usable and Useful Data Downloads Under GDPR/CCPA Access Rights via Co-Design

Sophie Veys, Daniel Serrano, Madison Stamos, Margot Herman,
Nathan Reitinger[†], Michelle L. Mazurek[†], Blase Ur
University of Chicago, [†]University of Maryland

Abstract

Data privacy regulations like GDPR and CCPA define a *right of access* empowering consumers to view the data companies store about them. Companies satisfy these requirements in part via *data downloads*, or downloadable archives containing this information. Data downloads vary in format, organization, comprehensiveness, and content. It is unknown, however, whether current data downloads actually achieve the transparency goals embodied by the right of access. In this paper, we report on the first exploration of the design of data downloads. Through 12 focus groups involving 42 participants, we gathered reactions to six companies' data downloads. Using co-design techniques, we solicited ideas for future data download designs, formats, and tools. Most participants indicated that current offerings need improvement to be useful, emphasizing the need for better filtration, visualization, and summarization to help them hone in on key information.

1 Introduction

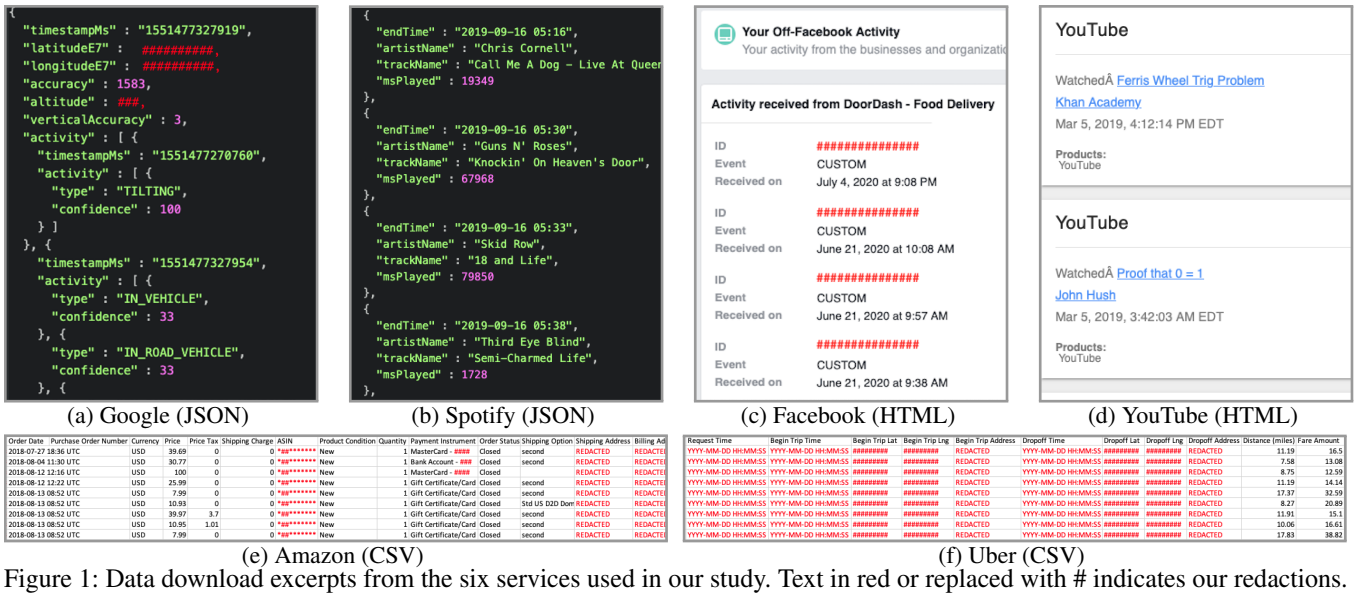
The principle of **data access** states that subjects should be able to obtain a copy of the data that has been collected about them. For decades, this principle has appeared in information privacy frameworks [24]. For example, access is one of the five core facets of the U.S. Federal Trade Commission's Fair Information Practice Principles (FIPPs) [24]. In past decades, while other FIPPs directly impacted consumers (e.g., the principle of notice underpins the ubiquity of privacy policies [66]), the principle of access was mostly ignored. In recent years, however, rights of access have been strengthened. In the European Union, Article 15 [79] of the General Data Protection Regulation (**GDPR**) enshrines a "right of access by the data subject." Similarly, under the California Consumer Privacy Act (**CCPA**), businesses must respond to consumer "requests to know" about data collected about them, enabling them " to access, view, and receive" a copy of that data [76].

Consumers might want access to their data for many reasons. First, data downloads can help users uncover distressing aspects of the online data ecosystem. Prior work has found that consumers can feel uneasy upon seeing evidence of online tracking and data collection [78, 81, 88]. Further, consumers often become upset when they feel that data has been misused or taken out of context [52], including for advertising [27] or politics [34]. In a widely discussed article, Hill used data downloads to expose "secret consumer scores" in which consumers' purchase histories and demographics impact their eligibility for refunds [31]. Access to data is a prerequisite for consumers to modify any incorrect information (the privacy principle of participation) [24]. Additionally, awareness of data collection might encourage users to exercise their right of erasure [9] or motivate other privacy-protective actions.

Privacy concerns aside, there are more practical reasons consumers might want access to their data. Many consumers have data spread across many platforms. For example, a consumer might have pictures published to Twitter, Instagram, and Tumblr. In the event they lose the device on which the original pictures are stored, they might try to reclaim as many photos as possible. Alternatively, a consumer might wish to move from one service (e.g., Spotify) to a competitor (e.g., Amazon Music), yet wish to seamlessly transfer their carefully curated playlists and other personal data. The pursuant right of **data portability**, which enables consumers to transfer personal data across services via interoperable formats, is also enshrined in both GDPR [79] and CCPA [76].

To comply with these legal rights of data access and portability, many companies have begun to offer what we term **data downloads**, which are either files or archives of files containing the identifiable data a business or other data processor has collected about a consumer. Figure 1 shows ex-

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.



(a) Google (JSON) (b) Spotify (JSON) (c) Facebook (HTML) (d) YouTube (HTML)

(e) Amazon (CSV) (f) Uber (CSV)

Figure 1: Data download excerpts from the six services used in our study. Text in red or replaced with # indicates our redactions.

ample excerpts from data downloads. While data downloads provide unprecedented access to the data companies hold about consumers, their format, organization, and design is highly variable across companies. As we discuss further in Sections 2–3, data downloads can range from individual CSV files to sprawling archives containing hundreds of gigabytes of data. In many cases, consumers are left to decipher files intended to be processed by computers. Many files are in JSON or CSV formats. They frequently use UNIX timestamps (see Figure 1a), rather than human-readable dates and times. Some data downloads even come as files containing a single line millions of characters long. Even archives in more typically human-readable formats like HTML can be disorganized and riddled with both jargon and undefined terminology. These sorts of problems led one journalist to subtitle an article about GDPR data downloads as "138GB of data and no real answers" [58]. Most data downloads appear intended to address both access and portability rights, arguably coming up short at providing humans meaningful transparency about their data.

Motivated by the unique opportunities data downloads afford, but also their apparent usability shortcomings, we asked:

- **RQ 1**: How do users react to both the format and content of their own data downloads?
- **RQ 2**: What information is important for users to see in their data downloads? What practical uses are imagined for this information?
- **RQ 3**: How should data downloads be redesigned to improve transparency and best support users' goals?

To address these questions, we conducted 12 online focus groups with a total of 42 participants. Each focus group centered on one of six companies offering data downloads: Amazon, Facebook, Google, Spotify, Uber, or YouTube. Participants came to the session with their own data download, which they had requested in a previous step of our protocol. In our sessions, participants were given time to explore their files, during which we gathered their opinions about both the format and content of current data downloads. Using co-design techniques, we then led participants through a series of activities designed to elicit their ideas and preferences for making data downloads more intelligible for humans.

Most participants indicated that current offerings need improvement to be usable and useful. Participants were generally unsatisfied with either the format or content of their data downloads, if not both. Despite the usability barriers of current formats, most discovered information they found surprising in their data download, commenting on the unexpected nature of information retained or the lengthy retention period. Participants emphasized the need for better filtration, interactive visualization, more meaningful organization, and summarizations that help them hone in on key information. Participants were also interested in understanding the contents of their data downloads at a higher level, including seeing aggregate statistics and how data is synthesized into inferences. Based on these findings, we offer recommendations for improving the presentation and intelligibility of data downloads.

2 Background and Related Work

We begin by highlighting the legal basis for data access rights. We then discuss prior research on those laws' impacts before focusing on prior work studying data downloads. We also summarize transparency efforts and co-design techniques.

Legal Basis for Access and Portability: The right of access and the right to data portability are provided under both GDPR [79] and CCPA [76]. The right of access mandates

that companies give consumers a full view of the data they hold about them upon request [4]. Both GDPR and CCPA prescribe the content that should be released, but not the format in which to release it. While GDPR Article 12 [79] requires the use of plain and clear language, the focus is on communication regarding the request, rather than the response itself [65]. Whether the response itself should be comprehensible is not fully specified, though mandating intelligible responses would be consistent with data access as a foundational privacy right. The right to data portability encompasses the transferability of data and enables users to change platforms, helping to prevent vendor lock-in. To support data portability, both laws do specify that data downloads should be readable by computers via standardized and interoperable formats [76, 79].

The rights of access and data portability may seem similar at first glance; both stipulate that data be made available to consumers. In fact, in CCPA the right to data portability is included within the right of access. However, this conflation of access and portability impairs comprehensibility. Machine readability and human readability are different standards requiring distinct approaches and mechanisms. We develop recommendations for human-intelligible data downloads.

Studies of GDPR and CCPA's Impacts: While we focus on data downloads under rights of access, prior work has explored other requirements and implications of GDPR and CCPA. Degeling et al. and Utz et al. studied cookie notices, finding a lack of usability in the consent process [18] and discussing the impacts on consumer choice [84]. Politou et al. studied the right to be forgotten and the right to withdraw consent, showing that the need to keep data for legal investigations may conflict with these rights [57]. Bertram et al. reported longitudinal data on how Google complied with those rights [9]. Biega et al. investigated the feasibility of data minimization, which demands that companies collect only the data necessary to satisfy the purpose of collection [11].

Researchers have also explored the effectiveness of the laws [26, 30, 32, 41, 83]. Mahieu et al. argued that GDPR is weakly enforced and would be more effective were it executed on a collective, rather than individual, level [44]. De Hert et al. found a range of interpretations for GDPR's data portability requirements, hypothesizing that data controllers might use formatting loopholes to prevent the full exercise of consumers' rights [30]. Grundstrom et al. [26] and Labadie [41] highlighted some compliance challenges data controllers face.

Data Downloads: While (to our knowledge) we are the first to study data downloads from a design perspective, others have investigated data downloads in other contexts. Martino et al. demonstrated that the right of access can be abused by using forged or publicly available data to make illegitimate data subject access requests (**DSARs**) for other people's data [45]. Boniface et al. also identified vulnerabilities in the authentication process and presented guidelines for improvement [14]. Bufalieri et al. [15], Urban et al. [82], Kröger et al. [40], and Spiller [74] made data download requests to data controllers, quantifying the response time [15, 82], evaluating the completeness of the data [15, 40], and documenting shortcomings in the request and authentication processes [15, 74]. Wei et al. had participants request their Twitter data, which they used to characterize ad targeting on Twitter and personalize a related user study [87]. Alizadeh et al. asked participants to request data downloads from loyalty card providers, interviewing participants about the request process and contents of their files [3]. Our work focuses on the design of data downloads themselves, as opposed to the request process.

Transparency Tools and Data Visualization: Although many users are concerned about their online privacy [20, 35, 46, 70], most do not understand important elements of the data-aggregation process [8, 36, 59, 61, 85, 90]. Profit motives tend to disincentivize full transparency [1, 47]. Researchers have attempted to provide additional transparency without platform support via black-box tools [5, 6, 17, 42]. Even with good intentions, conveying complex technical information to users is challenging [19, 21, 23, 71]. Many researchers have created transparency- and privacy-enhancing tools (**TETs** and **PETs**), such as browser extensions and dashboards [7, 8, 12, 13, 37–39, 43, 49, 49, 51, 55, 56, 62, 64, 67–69, 80, 89, 91]. Some tools highlight the need for effective visualizations in improving user understanding [7, 88]. Researchers also emphasize the need to provide users with direct, fine-grained control [12, 16, 38, 49, 55, 56, 62, 89]. Others argue that focusing on control over personal information unduly burdens users [28, 53, 66].

We take the first step toward the creation of GDPR/CCPA data download TETs and PETs via co-design sessions. Most prior work pre-dates or is unrelated to GDPR/CCPA data downloads, instead focusing on visualizing the types of information readily available to consumers at the time those tools were created. Datta found that dashboards show only a portion of the existing data [17]. We do not pre-select the information we deem interesting, such as inferences [60] or advertising [87]. Instead, we ask participants to highlight their desired content. While we confirm some best practices of data visualization generally [29, 72, 73], our recommendations are specific to the data-collection ecosystem and emerge from co-design based on participants' actual data downloads.

Co-Design: The co-design research method (sometimes called participatory design) includes end users in the design process to leverage the knowledge and skills of end users in collaboration with the expertise of researchers and designers [75]. Co-design has been used in a few prior studies of security and privacy tools [25, 50, 63, 86]. Weber et al. emphasized the importance of establishing a "common language" between participants and researchers [86]. Our own sessions build on the lessons of these prior applications of co-design.

3 Selection and Overview of Data Downloads

To facilitate concrete discussions, we centered each focus group on participants' own data downloads from one of six companies. Here, we explain how we chose those six companies and briefly describe their data downloads for context.

Company Selection: To select popular companies, we examined the privacy policies of the Moz Top 500 Websites [48] to see which let users download their data. We excluded 105 websites that were not in English, illegal (e.g., ThePirateBay), potentially embarrassing, or were unable to be accessed by the researchers. We then filtered for companies that allowed non-California and non-EU residents to make data subject access requests, resulting in 109 websites. We categorized each site using the Alexa Top 500 categories [2], assigning categories based on the service the company provides and the type of data expected to be found in its data download.

We selected companies based on the following criteria:

- A simple request process via a clear, online portal (no emailing, mailing, or calling required)
- Relatively quick fulfillment (less than 10 days when members of the research team requested their own data)
- An easily recognizable and popular company, making it easier to find participants with an active account
- Belonging to a category of company participants would likely use (e.g., social media, entertainment)

Further, we selected companies meeting the above criteria such that the final slate would encompass both the breadth and depth of file information potentially available across all data downloads, ensuring a reasonably representative sample of the types of information available. We chose the following six companies: **Amazon**[1] (shopping); **Facebook** (social media); **Google**[2] (location and search); **Spotify** (entertainment); **Uber** (transportation); and **YouTube** (media).

Data Downloads' Characteristics: Members of the research team requested their own data downloads for these companies and many others, recording the data types and corresponding format of each type of information available. This achieved three goals: (i) it informed us about the types of data available in each download; (ii) it enabled us to identify variation in data formats (e.g., UNIX vs. UTC timestamps); and (iii) it provided us with an initial impression of how human-readable each download was. Although we were able to interpret most of the available data, there were a number of items we could not resolve. For example, we were unable to interpret Uber's "horizontal accuracy" column, which contained values like "30" and "10."

[1] While Amazon offers data downloads for all products, including Kindle and Audible, we omit all but order history to keep sessions focused.

[2] Similarly, Google downloads can be very large and variable, so we omit all Google products except location and search.

Table 1 summarizes key aspects of team members' data downloads from these six companies; results for others may vary. Note that user-uploaded files, such as Facebook photos, YouTube videos, and Google Drive files, retain their original file format in the data downloads and are excluded from Table 1. Informally analyzing our own data downloads, we identified eight classes of information. We found extensive variation in how different companies included and presented data within these classes. For example, five companies' data downloads (all but YouTube) contained some sort of location data. Spotify's location data included the user's address, payment country, payment card postal code, family plan address, and Car Thing accessory shipping address. Facebook, in contrast, included the user's primary location, the current city included in their profile, the IP addresses and locations from which they had ever logged in, and the places where the user had checked in. Google's location data included time-stamped locations, data presumably collected from a phone GPS (latitudes, longitudes, velocities, altitudes), and the type of activity performed at a location. Appendix C gives other examples. These examples are intended to provide context for participants' comments, rather than being exhaustive.

4 Co-Design Study: Method

We conducted a three-part study than ran from July 2020 to September 2020. Part 1 was a screening survey. Eligible participants were asked to download their data from one of the aforementioned six companies and were invited to Part 2. Part 2 determined eligibility for Part 3, a 75-minute co-design session hosted on Google Meet. We recruited participants on Prolific, a crowdsourcing platform that has many advantages [54] over Amazon Mechanical Turk. The appendix contains the text of all survey instruments and focus group guides.

4.1 Participant Selection

In Part 1, participants completed a demographic and screening survey in Qualtrics to provide information that would help us create the co-design sessions. Participants indicated their availability for a focus group and chose the companies on our list of six for which they had active accounts. We compensated $1 USD for this survey, which took on average 2.5 minutes.

Based on the Part 1 responses, we selected prospective participants. We assigned one of the six companies to each participant. In Part 2, we provided participants instructions to request a data download from their assigned company. For companies that provided both HTML and JSON options (see Table 1), we instructed participants to select HTML as it is more likely to be human-intelligible. Google offered location data in both JSON and KML, but we opted for JSON due to the complexity of opening a KML file. Participants completed a second survey to verify they had successfully requested their

Table 1: Key aspects of data downloads, as obtained by the research team. For Amazon and Google, we report on both the subset (*) of the download used in our study and the full (FULL) versions of these data downloads.

Company	File Formats	Includes "ReadMe"?	Time of Receipt	# Folders	# Files	Size
Amazon*	CSV	No	2-5 days	<10	<10	KBs
Amazon FULL	CSV	No	4 weeks	Tens	Tens	MBs
Facebook	JSON or HTML	Yes	Almost instantaneous	Hundreds	Thousands	GBs
Google*	JSON, HTML, KML	Yes	Almost instantaneous	<10	<10	KBs
Google FULL	JSON, HTML, KTML	Yes	Hours	Tens	Hundreds	GBs
Spotify	JSON	Yes	1-2 weeks	<10	<10	KBs
Uber	CSV	Yes	Hours	<10	<10	KBs
YouTube	JSON, HTML	Yes	Almost instantaneous	<10	10–20	KBs

data download by pasting in text (with no identifying information) from the notification email or data download page. The Part 2 survey also asked about their general sentiments toward data access and privacy. Participants were compensated $2 for completing Part 2, which took 11 minutes on average.

Participants who completed Part 2 were invited to Part 3, a 75-minute focus group and co-design session centering on the company for which they had downloaded their data. Group sizes ranged from 3–5 based on participant availability and turnout. We ensured there were no more than two participants per session who were students, and no more than one participant per session with CS or IT expertise. We held two focus groups for each of the six companies, resulting in 12 focus groups in total. Participants were compensated $25 for Part 3.

Due to COVID-19, we held all focus groups remotely as video calls, recording only the audio. We used Google Meet because it provides real-time captioning (transcription). A researcher listened to all audio recordings and corrected the transcripts. Meet requires participants to log in with a Google account, displaying the associated name on-screen. To protect participant privacy, we made five anonymous Google accounts for participants to use. We turned off those accounts' activity tracking and ad personalization. We logged participants out and changed the passwords between sessions.

4.2 Structure of Focus Group Sessions

Each 75-minute session included several activities designed to encourage discussion and inspire ideas about improving data downloads. A third survey was conducted concurrently with the session to facilitate giving participants instructions and collecting written responses. We iterated on the design of our focus group protocol through five pilot sessions with convenience samples. After each, we incorporated feedback from the previous pilot session to clarify the wording of questions and instructions, correct typos, and improve logistics. As suggested by a pilot tester, in our final protocol we screen-shared slides with bullet-point instructions. The survey and slides aimed to help participants stay on track even if they experienced connectivity issues or other interruptions.

Introductions and Guidelines: We began by directing participants to the third survey in order to consent to both participation and audio recording. To help participants get to know one another, we asked participants to introduce themselves with their first name (real or fake) and a fun fact about themselves. In the chat window, we mapped anonymous Google account names (e.g., Participant 1) to the first name provided by the participant, allowing participants to refer to each other by name during discussions. We then gave participants general instructions for the session: to turn cameras on (a requirement aimed to increase engagement), to mute when they were not speaking, and not to take screenshots or make recordings. We reminded participants they were not required to share specific information about themselves or their data.

GDPR/CCPA 101 and Free Exploration: The first activity, intended to provide context about data downloads, was a minute-long overview of GDPR and CCPA. We explained that data downloads are available in part as a right granted to residents of the EU and California. We answered (to the best of our knowledge) any questions participants posed about these laws. We then had participants freely explore their data download for five minutes. We asked participants to inspect the format, content, and organization of the files. For sites with Read Me or HTML overviews (all but Amazon), we gave participants 1–2 minutes to read them. We encouraged them to comment aloud about anything they found interesting.

Scavenger Hunt and Discussion: While free exploration avoids priming participants about what to look at, it can also lead to a lack of engagement. Thus, we next asked participants to complete a scavenger hunt with their data downloads. We provided a list of items to find, such as a deleted message or the timestamp of a purchase. We selected items such that:

- Collectively, the items spanned multiple folders
- Items were not too difficult to interpret
- Some items might interest the participant (e.g., "What 'life stage' does Facebook assign to your friends?")
- Items required scrolling (e.g., "Find an album... that starts with the same letter as your first name.")
- Some items required cross-referencing multiple files

All scavenger hunt items, 6–11 per company, met at least one of these criteria. While the scavenger hunt inevitably introduced some bias, we believe these items helped to expose participants to a broad range of their data, including information they may have overlooked during free exploration. The scavenger hunt lasted 5–7 minutes. For privacy, participants did not enter their answers in the survey, nor read them aloud.

For 10–15 minutes, participants then discussed their first impressions of data downloads. We debriefed the scavenger hunt, asking about experiences navigating the data download and looking for items. We asked about the content and format of these files, as well as data-collection practices in general.

Highlight Activity: In our Qualtrics survey, participants were then given a list of folder names associated with the relevant company's data download and were asked to highlight the categories they would be most interested in seeing. This was designed to identify content participants cared about. There was no limit on the number of items they highlighted.

Data Viz 101: To inform and inspire participants, we held a five-minute introduction to data visualization. We asked participants to browse Information Is Beautiful [33], which visualizes daily news. We chose this site because it offered many options, rather than endorsing one or two specific visualization approaches. We also wanted to avoid visualizations that would alienate participants. Instead, we wanted them to focus on data presentation, rather than content. We asked participants to share examples of visualizations they found particularly interesting or well-designed, as well as examples that synthesized multiple pieces of information. Participants pasted links to visualizations, briefly summarizing what they liked about each. We then used a basic example to show that even simple visualizations can be effective, showing a spreadsheet with two columns ("month" and "number of cats petted") and graphing the data as a line chart.

Sketch Activity: Finally, we asked participants to sketch, either on paper or digitally, their ideal version of a visualization tool for their data download. All prior activities were designed to build up to this activity, which directly supported our ultimate goal: to work with participants to reimagine data downloads. We provided guiding questions, referencing content, formatting, and menu options. We told participants they could use any approach they wanted, but mentioned two possible options: a high-level approach sketching the general layout of a tool and specifying its different options, and a low-level approach focused on representing a specific type of information (e.g., location data). After uploading their sketches to our server, participants were asked to explain them. With participants' permission, we screen-shared their sketches to the group; we have also made them available for download [77]. Participants were then redirected to Prolific for compensation.

4.3 Data Analysis

We analyzed the data from our co-design sessions using affinity diagramming, a method for consolidating qualitative data into emergent groups or themes [10]. Two researchers used Miro, an online whiteboard, to collaboratively affinity-diagram comments from all 12 sessions. We placed meaningful quotes from all the session activities on virtual Post-it Notes, then grouped them with other similar quotes. We determined meaningful quotes to be everything that was shared during a session with the exception of what researchers said and moments when participants required clarification or when they experienced technical difficulties. We framed our groupings around "what" (data content) and "how" (data format). We then isolated themes within those top-level groupings. As needed, we split quotes to ensure they did not contain more than one cohesive idea. If a quote fit into more than one grouping, we duplicated it as needed.

We analyzed all quotes using pseudonyms containing an abbreviation for the company under discussion, Session A or Session B for that company, and an assigned participant number (from 1 to 5) during the session. An example pseudonym is *G-A-1*: the first participant in Session A for Google.

4.4 Protection of Participants

Our protocol was reviewed by the University of Chicago IRB and determined to be exempt. We collected no personally identifiable information. Study-related communication was conducted via Prolific's internal messaging system, which uses pseudonyms to identify participants. As discussed above, we did not ask participants to share their data downloads with us, we created anonymous Google accounts to avoid participants exposing their personal information, and during the session participants identified themselves using only their first name or a pseudonym. We did not video record the session or take screenshots, and we instructed participants not to do so either. Participants consented to audio recording of sessions before completing Part 1 and again before completing Part 3. We reminded participants at the beginning of the session that they were under no obligation to share specific information about themselves or their data. We also told participants that if they said something they did not want on record, they could let us know afterwards and we would delete that portion.

4.5 Limitations

Due to the rich qualitative nature of our study, we had a relatively small sample size (42 participants). We recruited only participants located in the U.S. As is typical on Prolific, our participants skewed younger and more educated than the average population of the U.S. Additionally, our study made technical demands of participants. They needed to download their data (aided by our instructions), join a Google Meet call (requiring a webcam and microphone), and upload a photo of

their sketch. These requirements were listed in our recruitment ad's eligibility section. As a result, it is likely that our sample excluded people with limited technological experience. Finally, our sample likely excluded those with disabilities, particularly visual impairment. Future work should investigate the accessibility of data downloads to those with disabilities.

As with any qualitative study, a participant not making or responding to a statement does not mean they disagree with it. While, for context, we provide counts of participants who expressed specific sentiments, we do not intend them to indicate overall prevalence. Additionally, the scavenger hunt activity may have primed participants. Though we tried to offset this concern by starting with a free exploration, it is possible that some participants may have been led to believe some sections of their data were most important based on our scavenger hunt items. As a result, we make no claims about the generalizability of our study. Rather, we present initial findings and directions for the design of data downloads.

5 Results

We first summarize participant demographics. We then report key findings from our focus groups in four key areas: reactions to existing content; ideas for improving content; reactions to existing formats; and ideas for improving formats.

5.1 Participants

We recruited 272 participants for Part 1, 77 of whom completed Part 2 and 42 of whom completed Part 3. Among Part 1 participants who did not continue, 156 never responded that they were ready for Part 2, while 39 were deemed ineligible.

Among Part 3 participants, 25 identified as male, 16 as female, and one as non-binary. Our sample skewed young: eight were 18–24, 19 were 25–34, nine were 35–44, five were 45–54, and one was 55 or older. Participants reported their highest level of educational attainment: two completed high school, nine completed some college or an associate's degree, 18 had a bachelor's degree, and 13 had a graduate or professional degree. Twenty-eight participants self-reported as White, seven as Asian or Pacific Islander, six as Black or African American, and one marked "other." Five indicated they were of Hispanic or Latino origin. Six had an education or were employed in computer science or IT. Six were students.

5.2 Content of Existing Data Downloads

We first report on participants' reactions to the content of their data downloads. Participants were struck by the heavy amount of detail, sometimes reaching the level of creepy, and found the inclusion of certain content surprising. They also identified several practical uses for their data downloads.

Expectations: Before each free exploration, we asked if participants had looked at their downloads before the session; only six of the 42 had. We then asked participants about their expectations of the content of their data downloads. **Most commonly, participants expected data downloads to contain demographics and data generated via interaction with the site** (e.g., friends and messages for Facebook, playlists and watch history for YouTube). Seven participants (all in Amazon, Facebook, or YouTube focus groups) also expected to see inferences or data from third parties.

During the free exploration and scavenger hunt, we encouraged participants to comment aloud. Nine were surprised by the presence or absence of information. For instance, F-B-3 was surprised to see facial recognition data, and S-A-1 was surprised to see her full address associated with a music company. F-B-2 and S-B-1, in contrast, expected to see more ad interests and search queries, respectively. For 12 participants (all companies), at least some of their expectations of what would be contained in their data downloads matched reality.

Twenty participants (all companies but YouTube) commented on the **accuracy or inaccuracy of their data**. Eleven participants mentioned that at least one part of their data was accurate, and 13 participants mentioned that at least one part was inaccurate. Seventeen participants (all companies) **mentioned that there was information missing from their data downloads**, either time gaps or information omitted altogether. Y-B-1 said, *"I personally think there's information that they have that's not in these files. And it can be used depending on what they need."* F-B-1 was surprised to find data he was *"100% sure"* he had erased. F-B-2 also found data he believed he had deleted, which he attributed to a *"legacy issue."* In contrast, Y-A-4 recalled deleting specific searches. He was surprised to find they did not appear in his data download. Regardless of whether these participants are correct about their deleted data, these comments suggest a lack of clarity and perhaps distrust related to how data is stored and retained. Y-B-1 commented, *"I do think there is information that might not be in these files, but somehow when we signed up for these platforms, in the very small fine text, they're letting us know it's there and you may not know what the term is or exactly what it means . . . but I do think there's other information that they capture that we may not be aware of."*

Reactions to Content: Fourteen participants (all companies) either commented on how far their data went back in time or reacted to old content. F-A-2 said, *"It's kind of weird to like be pulled back into that space of when you set up the first ever Facebook account."* Eight participants (all companies but Facebook) noted the **level of detail** in the files. Five were surprised by how detailed the files were; one partially blamed difficulty navigating the data on the level of detail. Eight participants (Facebook, Google, YouTube) reported feeling **creeped out or scared** about the breadth, detail, and type of data being collected and stored about them. Feelings of

unease were not always related to a lack of awareness. G-A-4 noted, *"For me, nothing was surprising. Like, I knew Google is recording everything. It's just that seeing this in front of me and all the data that has been collected over all the years, it's like a rude realization that yeah, there is someone watching you all the time."* G-B-2 also expressed this sentiment. This quotation highlights the potential for data downloads to be used for promoting privacy-protective behaviors. While many users are aware of tracking and data collection in general, a data download situates these practices in a personal context. This personal context is perhaps more likely to inspire action than simply hearing about data collection in the abstract.

Five participants (Amazon, YouTube, Uber) commented on the large size of their data download. G-A-3 noted the presence of many product options on Google's data download page. G-B-3 made a similar point: *"There was also so many other things you could download, that also really scared me. I was like, this is only my location and search history. I can't imagine if everything else was included."* A-B-2 felt the lack of definitions for terms made navigation and comprehension harder. Asked about navigating data downloads, F-B-1 said, *"It was like reading a book about myself but not written by myself."* This quotation is perhaps emblematic of a lack of control by users over their data. We note that the right to be forgotten, the right to participation, and the right to rectification can help users reclaim control over their data.

Uses and Misuses: Twenty-nine participants spanning all companies discussed possible uses or misuses of data downloads. Eleven (all but Facebook) identified **practical reasons why they might want access to their own data downloads**, including accessing a lost record, budgeting, or finding and erasing problematic information. Eight participants (Amazon, Facebook, Spotify, Uber) imagined these files could be used for **privacy purposes**, namely keeping track of the collection of their personal information. U-A-1 observed, *"It's interesting to understand how exposed you are from a privacy perspective."* Four participants (Facebook and Spotify) went beyond awareness, suggesting privacy-protective actions they or others might take after viewing their data downloads. F-B-1 mentioned refraining from making sensitive searches on Facebook, and S-A-2 and S-A-3 considered using information from a data download to help them keep their accounts secure. S-B-5 said, *"I think if the company had a data breach, and I knew that I was in that data breach, being able to see what data was potentially accessed for myself is important . . . if it has some kind of impact on my credit or I need to freeze my credit."* These privacy and security concerns arose organically from looking at the data, without prompting from researchers.

Mentioned misuses included account compromise or inappropriate targeted ads (six participants). Four mentioned **data downloads being used by law enforcement, the government, or in court**; three others agreed with or commented on these statements. F-A-1 said, *"I'm curious to know if this information can be subpoenaed in a court, because there's a lot of information here. So I mean if there's any illegal activity going on you could definitely use this file to find out."* Concerns about misuse of downloads themselves are not entirely misplaced [15,45]. We note that while law enforcement could likely obtain information directly from companies regardless of the data download feature, the data download did raise awareness of how much information companies store.

5.3 Desired Content

To help focus the efforts of programmers and designers who might craft interfaces for data downloads, we examined the types of data most and least interesting to participants. Participants discussed demographics, data generated via interaction with the company, inferences, and aggregate information.

Demographics and Site Data: Fourteen participants (all companies but Spotify) were interested in data associated with their demographics and direct site interaction, as opposed to inferences. Participants mentioned search history (three, Amazon and Facebook), location data (two, Facebook and Google), and photos (two, Facebook). Y-B-3 wanted to see *"how much personal information they've collected."* Y-B-1 agreed. Six Spotify participants also cited personally identifiable information and payment details as among the most important things to see. Four participants (Amazon, Uber, Facebook) wanted to know how their information was being used and shared, and three participants (Google, Uber, YouTube) wanted to see everything the company had about them.

Inferences and Advertising: Ten participants (all but Uber) wanted to see inferences made about them or obtain insight into inferencing algorithms. F-A-2 said, *"What's interesting to me is how my online behavior is affecting how this company and all the affiliates see me. And in what category, say, they put me or don't put me. . . . That has a way broader implication than the actual things that I am looking for. . . . Who is programming these algorithms? . . . Do they represent a broader part of society or are they all from a very similar group, similar life experiences and backgrounds?"* Four participants (Amazon and YouTube) wanted insight into the company's recommendation algorithm and/or the data that powers it.

Seven participants (all but Spotify) wanted to see **advertising data**. Two (Facebook, YouTube) wanted data on advertising-related inferences. Furthermore, two (Google, Uber) mentioned things they said aloud being used for advertising, referencing a common folk belief that devices secretly listen to users [22]. U-B-2 said, *"There's nothing weirder than having talked to someone on the phone . . . and an ad pops up for something that you were talking to somebody on the phone about."*

Nine participants discussed things they did not want to see. Four (Amazon and YouTube) reported no interest in any of the information in their data downloads. Four others (Facebook, Spotify) named specific data types they found useless or irrelevant, including poke history, past aliases, or search queries. Three participants (Amazon, Facebook) wanted **less data retention, for privacy and security** purposes. For instance, F-B-3 did not want long-term location data saved because she feared that a malicious actor could use it to find her.

Aggregation and Synthesis: Participants wanted more than just raw data.[3] Nineteen participants (all but Facebook; seven unprompted) wanted to see aggregate data about their site usage or activity. Eight (Amazon, Uber, YouTube; two unprompted) mentioned wanting aggregate financial data for business, budgeting, or to examine spending habits. U-B-5, who drives for Uber, imagined using aggregate data to determine the most profitable times for her to drive. Six participants (Google, Spotify, YouTube) wanted a breakdown of how much time (relative or absolute) they spent listening to songs or artists, watching videos, or otherwise engaging.

During Data Visualization 101, we asked participants to look for examples synthesizing multiple data types. Participants quickly adopted this theme: 30 (all companies) used some kind of synthesis in their sketches. In addition, three participants (Amazon, Facebook, Spotify) brought up data synthesis, unprompted, earlier in the discussion. Most imagined using synthesis to learn about themselves and their use of the site. For example, S-B-5 proposed a graph of the correlation between music choices and the time of day and year.

5.4 Format of Existing Downloads

Participants identified benefits and drawbacks to the format and organization of the data downloads they examined.

Quantity of Data: Nine participants agreed that **accessing records was difficult due to the vastness of their data downloads**. Four (Facebook, Google, YouTube) described it as *"overwhelming,"* and five (Amazon, Facebook, Google, YouTube) described the challenge of moving through their data as *"tough," "tedious," "hard,"* or *"time-consuming."*

Navigation: Twenty-one participants (all but Amazon) felt that, **overall, their data download was easy to explore**. They attributed this to intuitive organization, nicely formatted files, and descriptive folder and file names. S-B-4 said, *"I think the names and the descriptions of the files was exactly what I expected them to be once you clicked on them."*

[3] We use "aggregate" for the collection of multiple instances of a single type of data (e.g., to summarize or identify trends). We use "synthesis" for combining multiple types of data to obtain insights unavailable in isolation.

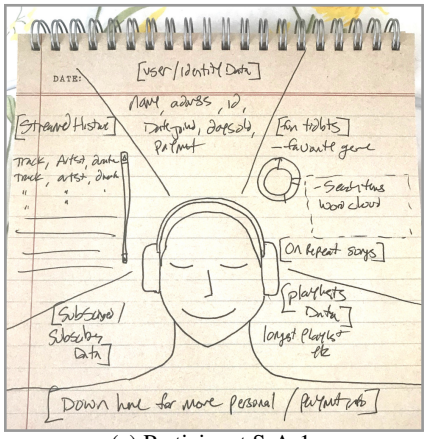
Conversely, 13 participants (all companies) expressed **difficulty navigating through their data downloads**. Nine of these (all but Spotify) attributed their difficulties to a **lack of familiarity** with data downloads in general, and with file types such as JSON specifically. Three of these nine said that, despite initial difficulties, they expected they could learn to navigate the files over time. U-B-2 described *"a very small learning curve where you had to figure out how the information was set up. . . . Once you figure that out, it's pretty easy."* Six participants felt they needed more than a single read-through to understand their files. A-B-4 said, *"Everything was the same font and the same size, so there's nothing bolded that will jump out at you."* Y-A-5 wondered about deliberate obfuscation: *"Most of the interesting data is stored in these files, that as a non-specialist, I can't read. . . . We're effectively illiterate when it comes to reading this additional data that they've been collecting."*

Organization: Thirteen participants spanning all companies felt the **files were disorganized** or could be more usefully organized. Eight participants (Amazon, Facebook, Spotify, YouTube) attributed their difficulty finding information to data downloads' disorganization. Y-A-1 said, *"The top-level organization makes a lot of sense, but then when you try and go one layer deeper then it just turns into raw data."* Y-A-3 remarked, *"It's like they didn't even try [to organize the data]. They just kind of dumped it on you."* Five participants (Amazon, Spotify, Uber) felt that **related information was inconveniently spread across multiple files**. A-A-4 said, *"I'd prefer it if it was just a single file,"* and A-A-1 agreed.

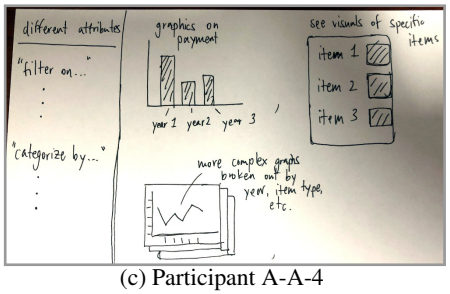
In contrast, nine participants (all but Google) were satisfied with the organization of their files. U-B-4 said, *"It looked exactly the way I would organize it."*Two commented on the **usefulness of folder names and how files were ordered**. S-B-5 noted that *"the 'follow list' was alphabetized, and then you could kind of see other stuff was by most recent."*

File Formats: Twelve participants (Google, Spotify, Uber, YouTube) discussed **difficulties with JSON files** and how they might deter others. G-B-3 said, *"A JSON file to begin with is pretty inaccessible."* S-A-1 said, *"I was kind of surprised that it . . . comes down in a JSON file, which I think could feel really intimidating."* Five participants (Google, Spotify, Uber, YouTube) **weren't able to open the JSON files** on their computers. On the other hand, S-B-5 found the JSON files *"user-friendly to see, especially with the way they color coded it."* Note that the color-coding was a feature of the JSON viewer we provided. G-A-4, the tech expert of his group, pointed out that JSON would enable analysis scripts.

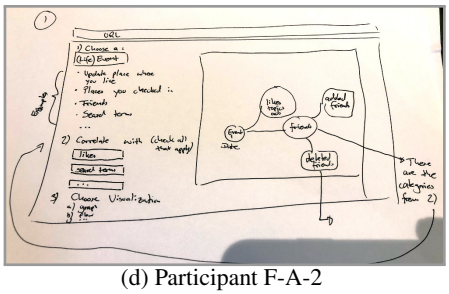
Three participants (Facebook and Google) felt **HTML files were usable and useful**, but not everyone agreed. Y-A-2 pointed out that an HTML file *"has a nice user interface and I can scroll through it all, but it's still not useful because*



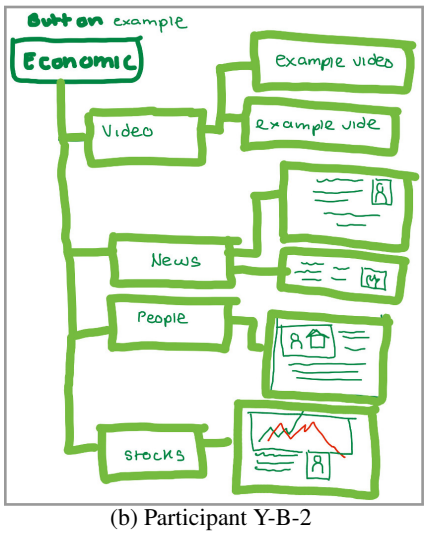
(a) Participant S-A-1



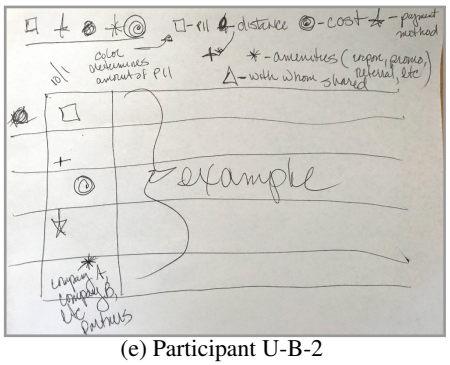
(c) Participant A-A-4



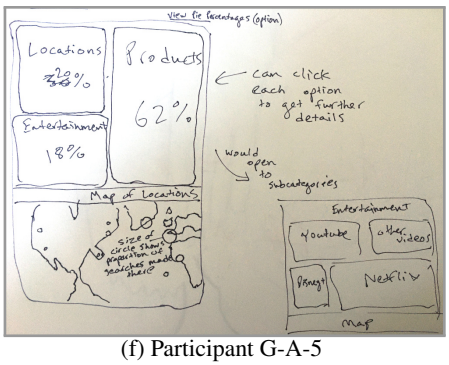
(d) Participant F-A-2



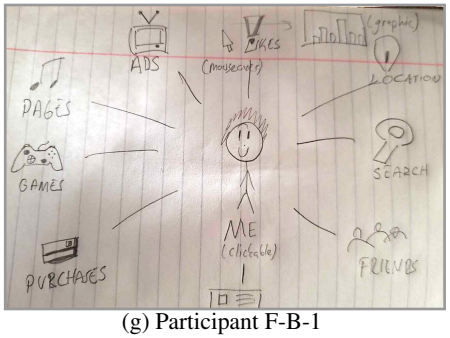
(b) Participant Y-B-2



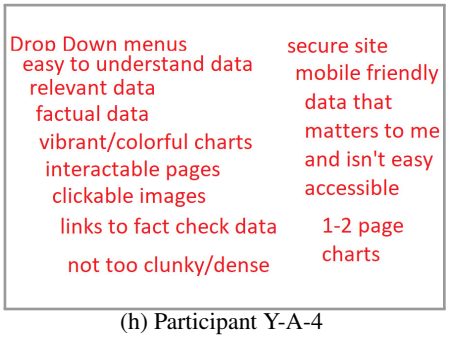
(e) Participant U-B-2



(f) Participant G-A-5



(g) Participant F-B-1



(h) Participant Y-A-4

Figure 2: Excerpts from participants' sketches during the design activity.

it's a long list to scroll through. If you spent more than a couple days on YouTube you can incur a very long list, and I've actually had that file crash multiple times." A-B-3 and U-A-1 found the CSV format of their data downloads straightforward, though A-B-3 added that she worked with spreadsheets daily. In contrast, A-A-4 did not find CSV files convenient.

5.5 Desired Format

Finally, we report on participants' ideas for formatting, including meaningful organization, filtration, visual representation, and interactivity. As in any co-design exercise, participants' suggestions should be seen as inspirations for design professionals to build on, rather than direct specifications.

File Formats and Interfaces: Participants suggested various improvements to the file formats. Y-A-2 liked that HTML files could be opened easily in the browser, but also wished for a better user interface. Several participants expressed interest in **CSV or other spreadsheet-compatible formats**. G-B-3, who had prior experience downloading CSV Twitter data outside this study, said the Google data would have been easier to digest had it been formatted similarly. G-B-2, Y-A-5, and A-A-1 also discussed the merits of spreadsheet formats.

G-B-2 considered printing out his data download, citing research that people comprehend information better when it's printed on paper. G-B-1 agreed, adding that an older generation might feel more comfortable with paper.

Finding Important Information: Twenty-one participants (all groups) expressed a desire for a **high-level overview of their data, with the option of delving for more information.** Participants wanted to see either an overview of everything contained in the download or a summary of the most important information. S-A-1 and F-B-1 used this approach in their sketches (Figures 2a and 2g). Y-A-5 said, *"The idea is to give them as much information as possible as an option, but not to overwhelm them with this sort of first glance, first blush dashboard."* This suggestion aligns with existing best practices in data visualization [29, 73].

Seventeen participants (all companies) emphasized **the importance of filtration**. F-A-1 commented, *"It would've been helpful to have filters so that you can organize the information by date or time, because if I'm looking for something specific, scrolling through that page of long history would be tedious."* F-A-2 agreed, and five other participants from Amazon and Google sessions made similar comments. Ten participants (all companies but Facebook) included filtration features in their sketches, including three who had already commented on filtration earlier in the session. Figure 2c shows A-A-4's sketch, which included a filtration feature.

Relatedly, 12 participants (all companies) included options to **sort by date** in their sketches. A-B-3 noted, *"I would only ever look at my stuff chronologically."* Most data downloads already organize relevant data by date. However, ten participants imagined extending this to filter by day, week, month, or year. F-A-2 imagined a different kind of sorting: an event-centered visualization in which the user selected a life event like *"[got] married, . . . moved to a new place, or got a new job."* Data would be displayed in relation to that event (Figure 2d). A-B-2 sketched another option: separating human-determined and computer-determined information.

In addition to sorting and filtering, several participants mentioned **prioritization**. Three Amazon participants mentioned reorganizing files so that meaningful information (e.g., item descriptions) appeared before less semantically useful data (e.g., order ID numbers). Three others (Facebook, YouTube) wanted to prioritize data types with the most entries.

Visualization: Nineteen participants (all companies) used **line graphs, bar graphs, pie charts, or tree graphs** in their sketches. Participants plotted information like payments vs. time, ads clicked on vs. ignored, and breakdowns of online activity by categories (e.g., entertainment, news, and people). Y-B-2's sketch (Figure 2b) illustrates the latter. This reflects participants' strong interest in synthesis, discussed above.

Nine participants' sketches (Amazon, Facebook, Google, Uber) used a **map to show location data**. Eight of these combined their map with other data, such as friends, search history, or frequency of the location. G-A-3 and S-A-3 used timelines in their sketches, consistent with the desire for chronological organization. Y-A-5 mentioned word clouds, and S-A-1 included a word cloud for search history in her sketch.

Fourteen participants (all companies) emphasized **interactivity** and/or included it when describing their sketches. Y-A-4, for example, wanted *"something that you'd be able to click on, whether you enlarge it or you control it with the mouse wheel . . . so you can see it a little more clearly or if you're looking for something specifically"* (Figure 2h). Others suggested buttons, menus, hovering, and clickable elements.

Several participants were also attracted to **simplicity** in visualization. Y-A-2 wanted to avoid *"very graphically interesting stuff that represents the data horribly. Because they don't do, like, bar graphs. They'll do, look at this zigzag graph, and you have no idea what that graph's trying to tell you or show you, because it doesn't actually tell you anything except for look pretty."* Five participants (Amazon and YouTube) emphasized simplicity in their sketch explanations.

Participants (19, all companies) also argued for using **color and element size** to distinguish data. Thirteen (all companies but Amazon) used color or size distinctions in their sketches. Six (Google, Spotify, Uber) used size and color to represent frequency, such as most-listened-to artists and most-visited locations. Figure 2f shows one example. These requests, which track good visualization practice, contrast heavily with the plain-text files participants viewed during the sessions.

Security, Privacy, and Accuracy: Four participants identified format-related security and privacy considerations. F-B-1 wished his data download had been **password-protected** so that someone with access to his computer could not read it, though presumably most information could also be accessed by visiting Facebook directly while logged in. Y-A-4 said he expected to access data via HTTPS and wished for some form of data encryption. The same participant also asked about fact-checking, or some other mechanism for verifying the provided data was accurate. G-B-2 and A-B-2 proposed in-band deletion of data directly after viewing it. This accords with the longstanding interaction principle of direct manipulation [72].

6 Discussion

We detail our design recommendations, followed by a brief discussion of the policy implications of our results.

6.1 Design Recommendations

Our co-design study provides insight into how to reimagine data downloads to be usable and useful. Our participants identified a variety of goals and use cases for data downloads. Some wanted easy access to artifacts, memories, or original content. Others wanted to know what personal information was collected and stored about them, or wanted insight into the inferences being made about them. A few were curious about aggregate statistics.

Participants also identified significant shortcomings of current data downloads that might hinder these goals: poor organization, sometimes-unfriendly file formats, and too many details with no way to filter. While these obstacles could perhaps be overcome with sufficient patience, they may deter users outside a research study. Furthermore, most data downloads do not include meaningful aggregation or synthesis, though it is highly likely such analysis is conducted internally to power recommendations and personalized advertising. The current state of data downloads thus prevents users from fully reaping the benefits that the right of access might provide.

It is therefore important to re-imagine data downloads for use by humans, separately from data designed for machine interpretation. Drawing from participants' responses, we make the following recommendations for data visualization tools:

- **Meaningful Organization:** Organize data chronologically, but with options for aggregating (e.g., by month). Group related data together. In line with visualization best practice, offer both a high-level overview and details on demand [73].
- **Filtration:** Enable filtering by type of data as well as other properties (e.g., payments over a certain amount).
- **Aggregation and Inferencing:** Provide insight into data aggregations and inferences made with the user's information, as well as the mechanisms behind them.
- **Interactivity and Exploration:** Enable rich interactions, such as selecting elements of high interest and zooming or hovering to reveal more information. Provide functions similar to simple spreadsheet or scripting tools to support synthesis. Current static formats (e.g., JSON, HTML) work with all common platforms, and interactive views should too. Web-browser-based interactivity may be appropriate.
- **Direct Manipulation:** Historically, rights of access have been associated with *participation*, the ability to contest or correct information [24]. Enact this principle by allowing correction via direct manipulation in the data download interface. Further, streamline the right to erasure (also defined in GDPR) via direct requests to delete specific information.

Efforts to improve data downloads could be made by the companies themselves, or by third parties. Companies know the most about their data, including its origin and schema, and are thus in the best position to provide explanations. Companies are also in the best position to enable direct manipulation for deleting or contesting data. However, companies may not have strong incentives to improve data downloads. From a legal compliance standpoint, data downloads are arguably sufficient currently. Additionally, some companies may wish to keep data downloads abstruse to hide unsavory data practices.

Thus, data downloads present an opportunity for third-party privacy and transparency advocates to design tools for user empowerment, continuing the mission of prior TETs and PETs. We found that viewing data downloads, even in their current not-very-usable state, raised organic privacy and security concerns. Third-party tool designers should consider how to make the content salient and digestible, while still leveraging the "creepy factor" [78, 88] to help users make better privacy choices. Additionally, third-party tools could offer cross-platform analysis and data-driven recommendations that promote privacy and support users in exercising control. Such tools could also serve as a GDPR/CCPA hub, allowing users to make data download and deletion requests with a click of a button. Further, third-party tools could (with proper consent and pseudonymization) aggregate data across users in order to characterize the data-collection ecosystem.

6.2 Policy Implications

The creation of data visualization tools should support, but not replace, legal intervention. While our sessions were designed to elicit ideas for data visualization tools, our findings also have implications for future iterations of data access laws:

- **Access vs. Portability:** We suggest that tensions between the right of access and right to data portability hinder the efficacy of the former. Future laws should better differentiate these requirements and include comprehensibility standards for data access rights. Specifically, data downloads should include a README-type file with an overview of the structure and content of the files, plus explanations for any technical or otherwise unintuitive fields.
- **Required Content:** We found that users were curious about their data, especially about inferences and aggregate data. It is at present ambiguous as to what data must be included in a data download. For instance, companies may argue that inference data embeds trade secrets and thus exclude inferences from data downloads. Notably, the Spotify and YouTube files did not include data about how recommended songs and videos were determined, which is related to the topic of algorithmic transparency. Policymakers should weigh companies' interests against users' right of access when deciding what data is within scope of a data subject access request. The law should clarify the required content.
- **Explanations for Missing Data:** Some participants felt data was missing from their files (e.g., gaps in time or the omission of certain categories). Data downloads should flag when and why data is missing.

The readability of data downloads is important not only for users, but also for technologists who will rely on README files with clear explanations to create data visualization tools. Thus, technology and law are both responsible for improving the transparency of the online data ecosystem.

7 Conclusion

We presented results from 12 focus groups with a total of 42 participants. We solicited participants' reactions to their own data downloads from one of six companies: Amazon, Facebook, Google, Spotify, Uber, or YouTube. Participants completed activities that familiarized them with their data downloads, elicited their opinions about content and format, and sparked inspiration for drawing their ideal data download visualization. Participants identified several key weaknesses in current data downloads, including that they were disorganized, unintuitive to navigate, and lacked usability features like filtration. These criticisms illuminate the need for companies themselves, or interested third parties, to reimagine data downloads to be usable and useful for humans, rather than simply machine-readable. This would better support the *right of data access*, as distinct from *data portability*. To this end, we presented associated design recommendations.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. CNS-2047827. We acknowledge funding from a UMIACS contract under the partnership between the University of Maryland and DoD. We thank Emma Veys, Joe Veys, Christopher John Boyle, Purrsephone, and Furdinand for their assistance. We also thank Lior Strahilevitz and the attendees of the 2021 Privacy Law Scholars Conference for their feedback and comments.

References

[1] Alessandro Acquisti, Curtis Taylor, and Liad Wagman. The Economics of Privacy. *Journal of Economic Literature*, 54(2):442–492, 2016.

[2] Alexa. The Top 500 Sites on the Web. https://www.alexa.com/topsites/category, 2020.

[3] Fatemeh Alizadeh, Timo Jakobi, Alexander Boden, Gunnar Stevens, and Jens Boldt. GDPR Reality Check – Claiming and Investigating Personally Identifiable Data from Companies. In *Proc. EuroUSEC*, 2020.

[4] David Alpert. Beyond Request-and-Response: Why Data Access will be Insufficient to Tame Big Tech. *Columbia Law Review*, 120:1215–1254, 2020.

[5] Athanasios Andreou, Márcio Silva, Fabrício Benevenuto, Oana Goga, Patrick Loiseau, and Alan Mislove. Measuring the Facebook Advertising Ecosystem. In *Proc. NDSS*, 2019.

[6] Athanasios Andreou, Giridhari Venkatadri, Oana Goga, Krishna Gummadi, Patrick Loiseau, and Alan Mislove. Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations. In *Proc. NDSS*, 2018.

[7] Julio Angulo, Simone Fischer-Hübner, Tobias Pulls, and Erik Wästlund. Usable Transparency with the Data Track: A Tool for Visualizing Data Disclosures. In *Proc. CHI*, 2015.

[8] Rebecca Balebako, Jaeyeon Jung, Wei Lu, Lorrie Faith Cranor, and Carolyn Nguyen. "Little Brothers Watching You:" Raising Awareness of Data Leaks on Smartphones. In *Proc. SOUPS*, 2013.

[9] Theo Bertram, Elie Bursztein, Stephanie Caro, Hubert Chao, Rutledge Chin Feman, Peter Fleischer, Albin Gustafsson, Jess Hemerly, Chris Hibbert, Luca Invernizzi, et al. Five Years of the Right to be Forgotten. In *Proc. CCS*, 2019.

[10] Hugh Beyer and Karen Holtzblatt. *Contextual Design: Defining Customer-Centered Systems*. Morgan Kaufmann, 1998.

[11] Asia J. Biega, Peter Potash, Hal Daumé III, Fernando Diaz, and Michèle Finck. Operationalizing the Legal Principle of Data Minimization for Personalization. In *Proc. SIGIR*, 2020.

[12] Christoph Bier, Kay Kühne, and Jürgen Beyerer. PrivacyInsight: The Next Generation Privacy Dashboard. In *Proc. APF*, 2016.

[13] Debmalya Biswas, Imad Aad, and Gian Paolo Perrucci. Privacy Panel: Usable and Quantifiable Mobile Privacy. In *Proc. ARES*, 2013.

[14] Coline Boniface, Imane Fouad, Nataliia Bielova, Cédric Lauradoux, and Cristiana Santos. Security Analysis of Subject Access Request Procedures How to Authenticate Data Subjects Safely When They Request for Their Data. In *Proc. APF*, 2019.

[15] Luca Bufalieri, Massimo La Morgia, Alessandro Mei, and Julinda Stefa. GDPR: When the Right to Access Personal Data Becomes a Threat. In *Proc. ICWS*, 2020.

[16] Johana Cabinakova, Christian Zimmermann, and Guenter Mueller. An Empirical Analysis of Privacy Dashboard Acceptance: The Google Case. In *Proc. ECIS*, 2016.

[17] Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. In *Proc. PETS*, 2015.

[18] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. We Value Your Privacy ...Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *Proc. NDSS*, 2019.

[19] Claire Dolin, Ben Weinshel, Shawn Shan, Chang Min Hahn, Euirim Choi, Michelle L. Mazurek, and Blase Ur. Unpacking Perceptions of Data-driven Inferences Underlying Online Targeting and Personalization. In *Proc. CHI*, 2018.

[20] Serge Egelman, Adrienne Porter Felt, and David Wagner. Choice Architecture and Smartphone Privacy: There's A Price for That. In *Proc. WEIS*, 2012.

[21] Motahhare Eslami, Sneha R. Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. Communicating Algorithmic Process in Online Behavioral Advertising. In *Proc. CHI*, 2018.

[22] Bree Fowler. Is Your Smartphone Secretly Listening to You? Consumer Reports, July 10, 2019. https://www.consumerreports.org/smartphones/is-your-smartphone-secretly-listening-to-you/.

[23] Fatih Gedikil, Dietmar Jannach, and Mouzhi Ge. How Should I Explain? A Comparison of Different Explanation Types for Recommender Systems. *International Journal of Human-Computer Studies*, 72(4):367–382, 2014.

[24] Robert Gellman. Fair Information Practices: A Basic History. *SSRN 2415020*, 2019. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2415020.

[25] Peter Leo Gorski, Yasemin Acar, Luigi Lo Iacono, and Sascha Fahl. Listen to Developers! A Participatory Design Study on Security Warnings for Cryptographic APIs. In *Proc. CHI*, 2020.

[26] Casandra Grundstorm, Karin Väyrynen, Netta Iivari, and Minna Isomursu. Making Sense of the General Data Protection Regulation—Four Categories of Personal Data Access Challenges. In *Proc. HICSS*, 2019.

[27] Julia Hanson, Miranda Wei, Sophie Veys, Matthew Kugler, Lior Strahilevitz, and Blase Ur. Taking Data Out of Context to Hyper-Personalize Ads: Crowdworkers' Privacy Perceptions and Decisions to Disclose Private Information. In *Proc. CHI*, 2020.

[28] Woodrow Hartzog. The Case Against Idealising Control. *European Data Protection Law Review*, 4:423, 2018.

[29] Jeffrey Heer and Ben Shneiderman. Interactive Dynamics for Visual Analysis. *Communications of the ACM*, 55(4):45–54, April 2012.

[30] Paul De Hert, Vagelis Papakonstantinou, Gianclaudio Malgieri, Laurent Beslay, and Ignacio Sanchez. The Right to Data Portability in the GDPR: Towards User-Centric Interoperability of Digital Services. *Computer Law and Security Review*, 2018.

[31] Kashmir Hill. I Got Access to My Secret Consumer Score. Now You Can Get Yours, Too. *The New York Times*, 2019. https://www.nytimes.com/2019/11/04/business/secret-consumer-score-access.html.

[32] Nicholas F. Palmieri III. Who Should Regulate Data?: An Analysis of the California Consumer Privacy Act and Its Effects on Nationwide Data Protection Laws. *Hastings Science and Technology Law Journal*, 2020.

[33] Information Is Beautiful, 2021. https://informationisbeautiful.net.

[34] Jim Isaak and Mina J. Hanna. User Data Privacy: Facebook, Cambridge Analytica, and Privacy Protection. *IEEE Computer*, 51(8):56–59, 2018.

[35] Ruogu Kang, Stephanie Brown, Laura Dabbish, and Sara Kiesler. Privacy Attitudes of Mechanical Turk Workers and the U.S. Public. In *Proc. USENIX Security*, 2014.

[36] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. "My Data Just Goes Everywhere:" User Mental Models of the Internet and Implications for Privacy and Security. In *Proc. SOUPS*, 2015.

[37] Farzaneh Karegar, Tobias Pulls, and Simone Fischer-Hübner. Visualizing Exports of Personal Data by Exercising the Right of Data Portability in the Data Track - Are People Ready for This? *IFIP International Summer School on Privacy and Identity Management*, pages 164–181, 2016.

[38] Patrick Gage Kelley, Paul Hankes Drielsma, Norman M. Sadeh, and Lorrie Cranor. User-Controllable Learning of Security and Privacy Policies. In *Proc. AISec*, 2008.

[39] Jan Kolter, Michael Netter, and Günther Pernul. Visualizing Past Personal Data Disclosures. In *Proc. ARES*, 2010.

[40] Jacob Leon Kröger, Jens Lindermann, and Dominik Herrmann. How do App Vendors Respond to Subject Access Requests? A Longitudinal Privacy Study on iOS and Android Apps. In *Proc. ARES*, 2020.

[41] Clément Labadie and Christine Legner. Understanding Data Protection Regulations from a Data Management Perspective: A Capability-Based Approach to EU-GDPR. In *Proc. Wirtschaftsinformatik*, 2019.

[42] Mathias Lécuyer, Guillaume Ducoffe, Francis Lan, Andrei Papancea, Theofilos Petsios, Riley Spahn, Augustin Chaintreau, and Roxana Geambasu. Xray: Enhancing the Web's Transparency with Differential Correlation. In *Proc. USENIX Security*, 2014.

[43] Candice Louw. Modeling Personally Identifiable Information Leakage that Occurs through the Use of Online Social Networks. Master's thesis, University of Johannesburg, 2015.

[44] René L. P. Mahieu and Jef Ausloos. Harnessing the Collective Potential of GDPR Access Rights: Towards an Ecology of Transparency. *Internet Policy Review*, July 2020.

[45] Mariano Di Martino, Pieter Robyns, Winnie Weyts, Peter Quax, Wim Lamotte, and Ken Andries. Personal Information Leakage by Abusing the GDPR "Right of Access". In *Proc. SOUPS*, 2019.

[46] Aleecia M. McDonald and Lorrie Faith Cranor. Beliefs and Behaviors: Internet Users' Understanding of Behavioral Advertising. In *Proc. TPRC*, 2010.

[47] Jeremy B. Merrill and Ariana Tobin. Facebook Moves to Block Ad Transparency Tools — Including Ours. ProPublica, January 28, 2019. https://www.propublica.org/article/facebook-blocks-ad-transparency-tools.

[48] Moz. The Moz Top 500 Websites, 2021. https://moz.com/top500.

[49] Min Y. Mun, Donnie H. Kim, Katie Shilton, Deborah L. Estrin, Mark H. Hansen, and Ramesh Govindan. PDVLoc: A Personal Data Vault for Controlled Location Data Sharing. *ACM Transactions on Sensor Networks*, 10(4), 2014.

[50] Cosmin Munteanu, Calvin Tennakoon, Jillian Garner, Alex Goel, Mabel Ho, Clare Shen, and Richard Windeyer. Improving Older Adults' Online Security: An Exercise in Participatory Design. In *Proc. SOUPS*, 2015.

[51] Patrick Murmann and Simone Fischer-Hübner. Tools for Achieving Usable Ex Post Transparency: A Survey. *IEEE Access*, 5, 2017.

[52] Helen Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2009.

[53] Ehimare Okoyomon, Nikita Samarin, Primal Wijesekera, Amit Elazari Bar On, Narseo Vallina-Rodriguez, Irwin Reyes, Álvaro Feal, and Serge Egelman. On the Ridiculousness of Notice and Consent: Contradictions in App Privacy Policies. In *Proc. ConPro*, 2019.

[54] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the Turk: Alternative Platforms for Crowdsourcing Behavioral Research. *Journal of Experimental Social Psychology*, 2017.

[55] Marta Piekarska, Yun Zhou, Dominik Strohmeier, and Alexander Raake. Because We Care: Privacy Dashboard on FirefoxOS. ArXiv, 2015.

[56] Marco Pistoia, Omer Tripp, Paolina Centonze, and Joseph W. Ligman. Labyrinth: Visually Configurable Data-Leakage Detection in Mobile Applications. In *Proc. MDM*, 2015.

[57] Eugenia Politou, Efthimios Alepis, and Constantinos Patsakis. Forgetting Personal Data and Revoking Consent Under the GDPR: Challenges and Proposed Solutions. *Journal of Cybersecurity*, 2018.

[58] Jon Porter. GDPR Makes It Easier to Get Your Data, but That Doesn't Mean You'll Understand It. The Verge, January 27, 2019. https://www.theverge.com/2019/1/27/18195630/gdpr-right-of-access-data-download-facebook-google-amazon-apple.

[59] Emilee Rader. Awareness of Behavioral Tracking and Information Privacy Concern in Facebook and Google. In *Proc. SOUPS*, 2014.

[60] Emilee Rader, Samantha Hautea, and Anjali Munasinghe. "I Have a Narrow Thought Process": Constraints on Explanations Connecting Inferences and Self-Perceptions. In *Proc. SOUPS*, 2020.

[61] Emilee Rader and Janine Slaker. The Importance of Visibility for Folk Theories of Sensor Data. In *Proc. SOUPS*, 2017.

[62] Philip Raschke, Axel Kupper, Olha Drozd, and Sabrina Kirrane. Designing a GDPR-compliant and Usable Privacy Dashboard. *IFIP International Summer School on Privacy and Identity Management*, 2017.

[63] Elissa M. Redmiles, Everest Liu, and Michelle L. Mazurek. You Want Me To Do What? A Design Study of Two-Factor Authentication Messages. In *Proc. SOUPS*, 2017.

[64] Robert W. Reeder, Patrick Gage Kelley, Aleecia M. McDonald, and Lorrie Faith Cranor. A User Study of the Expandable Grid Applied to P3P Privacy Policy Visualization. In *Proc. WPES*, 2008.

[65] General Data Protection Regulation. Guidelines on Transparency under Regulation 2016/679. 2018. https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=622227.

[66] Joel R Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T Graves, Fei Liu, Aleecia McDonald, Thomas B Norton, and Rohan Ramanath. Disagreeable Privacy policies: Mismatches between Meaning and Users' Understanding. *Berkeley Tech. LJ*, 30:39, 2015.

[67] Christopher J. Riederer, Daniel Echickson, Stephanie Huang, and Augustin Chaintreau. FindYou: A Personal Location Privacy Auditing Tool. In *Proc. WWW*, 2016.

[68] Florian Schaub, Aditya Marella, Pranshu Kalvani, Blase Ur, Chao Pan, Emily Forney, and Lorrie Faith Cranor. Watching Them Watching Me: Browser Extensions' Impact on User Privacy Awareness and Concern. In *Proc. USEC*, 2016.

[69] Roman Schlegel, Apu Kapadia, and Adam J. Lee. Eyeing Your Exposure: Quantifying and Controlling Information Sharing for Improved Privacy. In *Proc. SOUPS*, 2011.

[70] Carina Paine Schofield, Ulf-Dietrich Reips, Stefan Stieger, Adam N Joinson, and Tom Buchanan. Internet Users' Perceptions of 'Privacy Concerns' and 'Privacy Actions'. *International Journal of Human-Computer Studies*, June 2007.

[71] Nick Seaver. Knowing Algorithms. *Media in Transition*, 8, 2013.

[72] Ben Shneiderman. Direct Manipulation: A Step beyond Programming Languages. In *Proc. CHI*, 1981.

[73] Ben Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proc. VL/HCC*, 1996.

[74] Keith Spiller. Experiences of Accessing CCTV Data: The Urban Topologies of Subject Access Requests. *Urban Studies*, 53(13), 2016.

[75] Clay Spinuzzi. The Methodology of Participatory Design. *Technical Communication*, 2005.

[76] State of California. California Consumer Privacy Act. 2018. https://oag.ca.gov/privacy/ccpa.

[77] Study Participants. Sketches, 2021. https://www.blaseur.com/papers/soups21-sketches.zip.

[78] Omer Tene and Jules Polonetsky. A Theory of Creepy: Technology, Privacy and Shifting Social Norms. *Yale JL & Tech.*, 16:59, 2013.

[79] The European Parliament and the Council of the European Union. Regulation (EU) 2016/679 (General Data Protection Regulation), 2016. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

[80] Slim Trabelsi and Jakub Sendor. Sticky Policies for Data Control in the Cloud. In *Proc. PST*, 2012.

[81] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising. In *Proc. SOUPS*, 2012.

[82] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. A Study on Subject Data Access in Online Advertising After the GDPR. In *Proc. DPM*, 2019.

[83] Tobias Urban, Dennis Tatang, Martin Degeling, Thorsten Holz, and Norbert Pohlmann. Measuring the Impact of the GDPR on Data Sharing in Ad Networks. In *Proc. ASIACCS*, 2020.

[84] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. (Un)informed Consent: Studying GDPR Consent Notices in the Field. In *Proc. CCS*, 2019.

[85] Jeffrey Warshaw, Nina Taft, and Allison Woodruff. Intuitions, Analytics, and Killing Ants: Inference Literacy of High School-educated Adults in the US. In *Proc. SOUPS*, 2016.

[86] Susanne Weber and Marian Harbach. Participatory Design for Security-Related User Interfaces. In *Proc. USEC*, 2015.

[87] Miranda Wei, Madison Stamos, Sophie Veys, Nathan Reitinger, Justin Goodman, Margot Herman, Dorota Filipczuk, Ben Weinshel, Michelle L. Mazurek, and Blase Ur. What Twitter Knows: Characterizing Ad Targeting Practices, User Perceptions, and Ad Explanations Through Users' Own Twitter Data. In *Proc. USENIX Security*, 2020.

[88] Ben Weinshel, Miranda Wei, Mainack Mondal, Euirim Choi, Shawn Shan, Claire Dolin, Michelle L. Mazurek, and Blase Ur. Oh, the Places You've Been! User Reactions to Longitudinal Transparency About Third-Party Web Tracking and Inferencing. In *Proc. CCS*, 2019.

[89] Zhi Xu and Sencun Zhu. SemaDroid: A Privacy-Aware Sensor Management Framework for Smartphones. In *Proc. CODASPY*, 2015.

[90] Yaxing Yao, Davide Lo Re, and Yang Wang. Folk Models of Online Behavioral Advertising. In *Proc. CSCW*, 2017.

[91] Angeliki Zavou, Vasilis Pappas, Vasileios P. Kemerlis, Michalis Polychronakis, Georgios Portokalidis, and Angelos D. Keromytis. Cloudopsy: An autopsy of data flows in the cloud. In *Proc. HCII*, 2013.

A Study Protocols

A.1 Consent Form (Shown Before Surveys in Parts 1 and 3)

Description: We are researchers at the University of Chicago doing a research study about data visualization. We hope to generate ideas for a data visualization tool based on your ideas and opinions. This is a three-part study.

- **Survey 1** – a short 5-minute screening survey.
- **Survey 2** – if selected, you will be asked to download your data from an online company. You will complete survey 2, a 10-minute survey in which you will verify that you have received your data and will choose a date and time for 75 minute online focus group with 2-4 other participants.
- **Survey 3 and Focus Group** – during this online session, we will lead you through a survey with a series of activities to inspire ideas for a tool that would visualize the data that you downloaded. These sessions will take place through Google Hangouts or a similar platform. The sessions will be audio-recorded.

Your participation is voluntary.

Incentives: You will receive $1 for completion of the first survey. You will receive $2 for completion of the second survey, which verifies that you have downloaded your data. You will receive $25 for completion of the third survey and participation in the video call.

Risks and Benefits: Your participation in this study does not involve any risks to you beyond those of everyday life. Taking part in this research study may not benefit you personally, but we may learn new things that could help others.

Confidentiality:

- No personally-identifiable information will be collected from you.
- If you decide to withdraw from this study, the researchers will ask you if the information already collected from you can be used.
- Any reports and presentations about the findings from this study will not include your name or any other information that could identify you. In some cases, you might provide personal stories or beliefs that we might quote or paraphrase as part of our research findings – any personally identifying information will be removed to protect your privacy.
- Identifiable data will never be shared outside the research team.
- De-identified information from this study may be used for future research studies or shared with other researchers for future research without your additional informed consent.

Contacts & Questions:

If you have questions or concerns about the study, you can contact Blase Ur, Assistant Professor, Department of Computer Science, University of Chicago. blase@uchicago.edu or (773)834-3034.

If you have any questions about your rights as a participant in this research, feel you have been harmed, or wish to discuss other study-related concerns with someone who is not part of the research team, you can contact the University of Chicago Social & Behavioral Sciences Institutional Review Board (IRB) Office by phone at (773) 702-2915, or by email at sbs-irb@uchicago.edu.

Consent:

Participation is voluntary. Refusal to participate or withdrawing from the research will involve no penalty or loss of benefits to which you might otherwise be entitled.

By clicking "Agree" below, you confirm that you have read the consent form, are at least 18 years old, and agree to participate in the research. Please print or save a copy of this page for your records.

◯ I agree to participate in the research
◯ I do NOT agree to participate in the research.

The transcriptions of the recordings taken as part of this research can be included in publications and presentations related to this research.

◯ Yes
◯ No.

A.2 Part 1 Survey (Demographics and Screening)

[Consent form]
Welcome to part 1 of the study. You will be asked a few demographic questions. If selected, you will be notified via Prolific with instructions for the next part.
What is your age? ◯ 18-24 ◯ 25-34 ◯ 35-44 ◯ 45-54 ◯ 55-64 ◯ 65 or older ◯ Prefer not to say
What is your gender? ◯ Male ◯ Female ◯ Non-binary ◯ Prefer to self-describe ◯ Prefer not to say
What is the highest degree or level of school you have completed? ◯ Some high school ◯ High school ◯ Some college ◯ Trade, technical, or vocational training ◯ Associate's degree ◯ Bachelor's degree ◯ Master's degree ◯ Professional degree or doctorate ◯ Prefer not to say
What is your race? Please select all that apply. ◯ White ◯ Black or African American ◯ American Indian or Alaska Native ◯ Asian or Pacific Islander ◯ Other (Please specify) ◯ Prefer not to say
Are you of Hispanic or Latino origin? ◯ Yes ◯ No
Which of the following best describes your educational background or job field? ◯ I have an education in, or work in, the field of computer science, computer engineering or IT. ◯ I do not have an education in, nor do I work in, the field of computer science, computer engineering or IT.
Are you currently a student? ◯ Yes ◯ No
As mentioned in the consent form and on Prolific, the final part of the study is a 75 minute online focus group for which you will be compensated $25. We will schedule this session based on your availability. Are you willing and able to participate in a Google Hangouts video call for this part of the study? ◯ Yes, I am willing to participate in a Google Hangouts call for the final part of the study. ◯ No, I am not interested in participating in the final part of the study.
Each of our focus groups will cover one of the sites below. We will use your answer to this question to place you into an appropriate group.

Please check all the sites for which the following is true:

1. You have an account.
2. You use the site frequently (at least once a month).
3. You have full ownership of the account. No one else has access.
4. You would be willing to download your data from this site. You will NOT be asked to send this data to us.

◯ Facebook ◯ YouTube ◯ Spotify ◯ Uber ◯ Amazon ◯ Google
When, in general, would you be available for a 75-minute video call? Please select all that apply. ◯ Monday morning ◯ Monday afternoon ◯ Monday evening

◯ Tuesday morning ◯ Tuesday afternoon ◯ Tuesday evening ◯ Wednesday morning ◯ Wednesday afternoon ◯ Wednesday evening ◯ Thursday morning ◯ Thursday afternoon ◯ Thursday evening ◯ Friday morning ◯ Friday afternoon ◯ Friday evening ◯ Saturday morning ◯ Saturday afternoon ◯ Saturday evening ◯ Sunday morning ◯ Sunday afternoon ◯ Sunday evening
Thank you for completing our screening survey. If selected, you will receive instructions for the next part on Prolific.

A.3 Part 2 Survey (Data Receipt and Knowledge)

Welcome to part 2 of the study. Today you will be asked to verify that you have downloaded and received your data. You will also be asked a few questions related to data, privacy, and the Internet. Finally, you will be asked to indicate your availability for a focus group session.
From which of the following companies did you request your data? This information can be found in your Prolific messages related to this study. ◯ Amazon ◯ Facebook ◯ Google ◯ Spotify ◯ Uber ◯ YouTube
[Participants were then asked to paste in text from an email related to their data download request or from the data download dashboard. We had company-specific screenshots and instructions to guide them. We did NOT ask them to provide any personally-identifiable information. We then asked these two questions:]
I have downloaded all the files that appear in the graphic above. ◯ Yes ◯ No
I know where these files are located on my computer. ◯ Yes ◯ No
Please write what you know about the General Data Protection Regulation (GDPR). Please do not look anything up. Your knowledge about this won't affect your eligibility or compensation in any way.
Please write what you know about the California Consumer Privacy Act (CCPA). Please do not look anything up. Your knowledge about this won't affect your eligibility or compensation in any way.
Which of the following terms have you heard of? Select all that apply. ◯ Data portability ◯ Right of access ◯ Right to be forgotten ◯ None of the above
Before this study, have you ever downloaded the data a company has collected about you? If so, which company?
Have you ever wanted to know what information a company has about you? ◯ Yes ◯ No ◯ Unsure
If companies gave you access to the information they had about you, what would you be most interested in seeing?
Have you ever been notified that data about you (passwords, emails, etc.) had been compromised? ◯ Yes ◯ No ◯ Unsure
What information, if any, do you think companies collect about you when you visit their sites?
Please follow this Doodle Poll link to schedule a time for part 3 of the study. The goal is to schedule the time that works for the most people.
Please observe the following guidelines:

- Enter your Prolific ID instead of your name
- Please choose ALL options that would work for you. This will increase your likelihood of being eligible to participate in part 3, a 75 minute focus group for which we offer compensation of $25.

Thank you for completing part 2 of the study. If you are eligible for part 3, based on your completion of this survey and your availability, we will message you on Prolific with more details.

A.4 Part 3 Survey (Focus Group)

Please do NOT start this survey until you have joined the Google Hangouts call. Check your Prolific inbox for information on how to join the call. Once you have joined, you may proceed to the next section.
[Consent form]
Please choose the company name designated on the PowerPoint. ◯ Amazon ◯ Facebook ◯ Google ◯ Spotify ◯ Uber ◯ YouTube
Please do not proceed to the next section until asked to do so by the session organizer.

Exploration of Files
Take 1-2 minutes to look at the index.html file. This is a visual overview of the folders and files contained in your data download. (Facebook sessions)
Take 1-2 minutes to look at the archive_browser.html file. This gives an overview of what is contained in the files, and also has links to settings related to your data. Make sure you look at all 3 tabs. (Google and YouTube sessions)
Take 1-2 minutes to look at the "Understanding My Data" link found in the Read Me First.pdf file. This gives an overview of what is contained in the files. (Spotify sessions)
Take 1-2 minutes to look at the readme.html file. This gives an overview of what is contained in the files, and also has links to settings related to your data. (Uber sessions)
Take 5 minutes to look through your data on your own. We encourage you to make comments aloud to us and to the other participants as you discover things that you find interesting.
While you look for these items, take time to familiarize yourself with your data, paying particular attention to the information that is included and the format and organization. Here are some things to thing about:

- What information is here?
- What information seems to be missing?
- How is this information presented?
- How is this information organized?
- How easy or difficult is it to find things that you are curious about?
- What, if anything, is confusing?

If you are unable to open your data file from your computer, use this online file viewer. Note: make sure you open this link in a **private browsing window.**
https://jsoneditoronline.org
To use this tool: [we included screenshots to supplement the written instructions]
Click the folder icon. Then click "Open from disk."
Find the data file you want to open and confirm.
You might want to open multiple tabs, one for each file in your data download. This way, you can refer back to a file without having to re-upload it.
Note: there is an option to save to cloud. To protect your privacy, do NOT use this feature.
Please do not proceed to the next section until asked to do so by the session organizer.

Scavenger Hunt

To get you acquainted with your data, we have a short scavenger hunt for you to complete. If you can't find an item, skip it and move on. The goal of this activity is to get you acquainted with your data.While some items can be easily found by looking on the website or app, please only look for the answers in the files that you downloaded. You are welcome to use Windows Explorer, Finder, or any other search tool on your computer. You may also use the [index.html, Read Me First pdf file and the "Understanding My Data" link, archive_browser.html, or readme.html] file found in your data download.
Again, please comment aloud as you find scavenger hunt items or anything else you find interesting.

[Below we have included the scavenger hunt items for all of the companies. The answers (in purple) were not displayed.]

Amazon

1. Have you ever used a gift card to make a purchase? In Retail.OrderHistory csv, "Payment Instrument Type" (column L), search for "Gift Certificate"
2. How many refunds have you been issued? In Retail.OrdersReturned csv OR Retail.CustomerReturns csv, count the number of rows excluding the top row
3. What was the reason for your most recent return? In Retail.CustomerReturns csv, "ReturnReason" (column F). Go to the last row to find the most recent.
4. Around what fraction of your orders are taxed? In Retail.OrderHistory csv, "Price Tax" (column G). Count all orders that have a non-zero tax value, then divide it by the total number of orders, which is the number of rows minus 1.
5. Around what fraction of your orders do you pay shipping? In Retail.OrderHistory csv, "Shipping Charge" (column G). Count all orders that have a non-zero tax value, then divide it by the total number of orders, which is the number of rows minus 1.
6. Around what fraction of your orders are sold directly from Amazon? In Retail.OrderHistory csv, "Marketplace" (column A). Count all orders sold by Amazon.com, then divide it by the total number of orders, which is the number of rows minus 1.
7. What's the date of your most expensive order (excluding tax and shipping charges) this year? In Retail.OrderHistory csv, first scroll through "Order Date" (column C) until you find orders from 2020. Then, look at "Price" (column F) until you find the most expensive order.
8. Find the product name of your most recently returned item. [Hint: this might require looking in more than one file.] First, you need to get the order ID. There are two ways to do this. (i) In Retail.OrdersReturned csv, find the last order ID (which is the most recent) in the orderID column (column C). (ii) In Retail.CustomerReturns csv, find the last order ID (which is the most recent) in the orderID column (column A). Now, search for that order ID number in Retail.OrderHistory. Once you've located the appropriate row, scroll over to find the product name (column Q).

Facebook

1. Which file contains Facebook search history? In search_history folder, your_search_history.html file
2. Find a friend request you sent. [Hint: you might want to check the Friends folder!] In friends folder, sent_friend_requests.html file
3. Find a Facebook user whose friend request you rejected or who you removed as a friend. In friends folder, rejected_friend_requests.html OR removed_friends.html files
4. Find the first documented Facebook page you liked. [Hint: your likes are stored chronologically in an html document, can you find it?] In likes_and_reactions folder, pages.html, last entry
5. What are some of your ad interests? Does anything surprise you? In ads_and_business folder, ads_interests.html
6. Find an advertiser who uploaded information about you. Do you recall ever interacting with that advertiser?In ads_and_business folder, advertisers_who_uploaded_a_contact_list_with_your_information.html
7. In approximately how many cities have you logged into Facebook? [Hint: that seems like it might be related to security!] In security_and_login_information, where_you're_logged_in
8. When did you register for your Facebook account? [Hint: it's not in the about you folder!] In profile_information folder, profile_information.html, value of Registration Date
9. How many events have you responded to in the past 6 months? In events folder, your_event_responses.html
10. What 'life stage' does Facebook think your friends are at? In about_you folder, friend_peer_group.html
11. What was the last date you updated your profile picture? In profile_information folder, profile_update_history.html

Google

1. What is the date of the most recent search in your history? In search folder, MyActivity.html
2. Find a search for a restaurant or business. In search folder, MyActivity.html
3. Find a search where you asked a question. In search folder, MyActivity.html
4. Find a search for a product you wanted to buy. In search folder, MyActivity.html
5. Find a search you made late at night. In search folder, MyActivity.html
6. Find a trip for which the mode of transportation was most likely a vehicle. In location history folder, Location_history.json, find high confidence number for vehicle
7. Find the latitude and longitude of a location that you likely traveled to on foot. In location history folder, Location_history.json, find high confidence number for on foot, then find corresponding lat/long pair
8. Find the specific address of a place you visited. In semantic location history folder, value of "address"

Spotify

1. How many users are you following? In follow.json, value of followingUsersCount
2. What is your display name? In identity.json, value of displayName
3. Find an album name from your library that starts with the same letter as your first name. If you can't find one, choose another letter. In yourlibrary.json, value of album
4. Find a search you made for a song or artist. In SerachQueries.json, value of typedQuery OR selectedQuery
5. According to your data, was your Spotify account created from Facebook? In Userdata.json, value of createdFromFacebook
6. Find the name of one of your playlists in your data. In Playlist1.json, value of name
7. What is the first song in the playlist you found for #6? In Playlist1.json, value trackName of first item
8. For how many milliseconds did you listen to the song you found in #7? [Hint: You might want to look at Streaming History.] If you've listened to it multiple times, pick one instance. In SteamingHistory0.json, search for the trackName found in #7, then it's the value of msPlayed
9. Find a song in your library where the track name is the same as the album name. In yourlibrary.json, value of album and value of track

Uber

1. What is the user rating associated with your account? In profile_data.csv, "Rating" (column E)
2. Where did you take your last recorded Uber from? In trips_data.csv, top row (most recent), "Begin Trip Address" (column H)
3. Were you referred to Uber? In profile_data.csv, "Referred to Uber?" (column J)
4. How many payment methods are listed under your account? In payment_methods, it's the number of rows minus 1
5. What is the longest (in terms of distance) Uber ride you've taken? In profile_data.csv, "Distance (miles)" (column m), find the largest
6. If you have an Uber Eats account–what was your last order and where was it from? [Hint: you might have to piece together this question from the available information!] Last order: eats_order_details.csv, "Item Name" (column F). Where it was from: eats_restaurant_names.csv, "Restaurant Name" (column C)

YouTube

1. Find a song you listened to on YouTube. In history folder, watch-history.html, scroll until you find a song
2. What is the date and time of the most recent video you watched on YouTube that was NOT music? In history folder, watch-history.html, most recent is at the top
3. Have you ever commented on a video? If so, find your oldest comment. In my-comments folder, my-comments.html, find the oldest one
4. Find a video you have watched that starts with the same letter as your first name. If you can't find one, pick another letter. In history folder, watch-history.html, scroll until you find a song
5. Find a search you made during a summer month. If you can't find one, pick another season. In history folder, search-history.html, scroll until you find a search
6. Have you ever uploaded a video to YouTube? If so, how many views did it get? If you have uploaded multiple, pick one. In videos folder, videoName.json, value of viewCount
7. Do you have any videos in your watch later list? If so, find the description of one of the videos in that list. In playlists folder, watch-later.json, value of description
8. Do you subscribe to any channels? If so, find the description of one of the channels you subscribe to. In subscriptions folder, subscriptions.json, value of description

Please do not proceed to the next section until asked to do so by the session organizer.

Highlight Activity

Amazon: Alexa; Amazon Drive; Amazon Music; Amazon Lists Wishlist; Amazon Smile Customer Data; Appstore; Customer Communication Experience; DSAR Customer Retail Addresses; Devices Registration; Digital Action Benefit; Digital Content Ownership; Digital Customer Attributes; Digital Prime Video Customer Title Relevance Recommendations; Digital Prime Video Location Data; Digital Prime Video View Counts; Digital Prime Video Viewing History; Kindle Reading Insights; Outbound Notifications Amazon Application Update History; Outbound Notifications Email Delivery Status Feedback; Outbound Notifications Notification Engagement Events; Outbound Notifications Push Sent Data; Outbound Notifications Sent Notifications; Payment Options Amazon Pay Browser Behavior Data; Payment Options Payment Instruments; Physical Stores Whole Foods; Prime Acquisition; Retail Amazon Custom; Retail Cart Items; Retail Customer Attributes; Retail Customer Contacts; Retail Customer Profile; Retail Customer Returns; Retail Customer Service Chats; Retail Gift Certificates; Retail Order History; Retail Orders Returned Payments; Retail Orders Returned; Retail Promotions; Retail Region Authority; Retail Reorder; Retail Sports Fan Experience; Retail Website Authentication Tokens; Search Data; Subscription and Digital Order History

Facebook: About You; Ads; Apps and Websites; Comments; Events; Followers and Following; Friends; Groups; Likes and Reactions; Location; Marketplace; Messages; Other Activity; Pages; Payment History; Photos and Videos; Posts; Profile Information; Saved Items and Collections; Search History

Google: Android Device Configuration Services; Arts & Culture; Calendar; Chrome; Classroom; Contacts; Crisis User Reports; Data Shared for Research; Drive; Fit; Fusion Tables; G Suite Marketplace; Google Help Communities; Google Input Tools; Google My Business; Google Pay; Google Photos; Google Play Books; Google Play Games Services; Google Play Movies & TV; Google Play Music; Google Play Store; Google Shopping; Google Translator Toolkit; Groups; Handsfree; Hangouts on Air; Home App; Keep; Location History; Mail; Maps; Maps (your places); My Activity; My Maps; News; Posts on Google; Profile; Purchases & Reservations; Reminders; Saved; Search Contributions; Shopping Lists; Street View; Tasks; Textcube; Voice; YouTube and YouTube Music; YouTube Gaming

Spotify: Car Thing; Family Plan; Follow; Identity; Payments; Playlist; Search Queries; Streaming History; User Data; Your Library

Uber: Account and Profile; Driver; Eats; Jump; Regional Information; Rider

YouTube: All Playlists; Likes; My Comments; Search History; Subscriptions; Uploads; Videos; Watch History; Watch Later

Please do not proceed to the next section until asked to do so by the session organizer.

Data Visualization 101

The ultimate goal of our project is to design tools that make it easier for you to understand your data downloads. One way to accomplish this is data visualization in which information is displayed using visuals like charts, graphs, and maps. Let's take a look at a couple of examples of data visualization.

When raw statistics or numbers are reported in the news, sometimes it can be hard to digest that information. Journalist David McCandless founded a site called informationisbeautiful.net, which offers visualizations of the daily news. Take a couple minutes to explore visualizations of the news you find most interesting. Open this link in a new tab: `https://informationisbeautiful.net`

Say I wanted to know how many cats I petted each month in 2019. One option would be to look at an excel spreadsheet with this information. However, if I wanted a visual representation of this data, I might take my spreadsheet and convert it to a line chart. I can easily look at this line chart and conclude that March was a great month for cat petting.

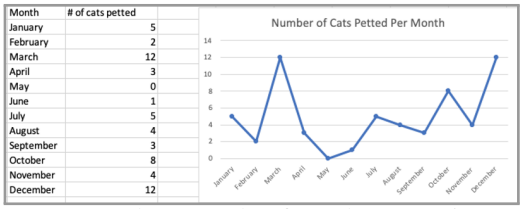


Figure 3: An example screenshot from the Data Viz 101 activity.

A.5 Focus Group Session Script

Hi. Thank you all for coming to this session. My name is ____ , and I will be co-leading this session with ____. Please begin the study on Prolific, which will take you to a Qualtrics survey. In part 1, you consented to participation in this study, which includes the audio recording of today's session. Please take a moment to confirm your consent to being recorded. Additionally, there will be a drawing activity. You will need a writing utensil and a piece of paper or the digital drawing tool of your choice. Please type "ready" in the chat once you have answered the consent questions and have a drawing tool on hand.

[Ensure that participants have completed consent form. Make sure everyone has pen and paper.]

Now, let's take a minute to introduce ourselves. Please say your first name, or the name by which you want to be referred during the session. Please also type your name in the chat when you are finished. For your protection, do not use your real last or middle name. Also, please share a non-sensitive fun fact about yourself. Finally, nominate someone to go next.

[One of the session leaders should go first to set the tone. "My name is ____ and I love cats. ____, go ahead!" Continue with introductions. We typed out a reference in the chat once everyone had introduced themselves: *Participant 1 - ____ Participant 2 - ____ . . . etc.*]

During today's session, you will be asked to look at your data and have the opportunity to answer questions aloud. Please remember that you are under no obligation to disclose specific information about you or your data during this session. We will also be recording the audio of this session. If you say something that you don't want on record, please let one of us know afterward, and we will delete that portion of the audio. We ask that everyone have their cameras turned on during the session. However, to protect your privacy, we will not record the video or take screenshots of the session. We ask that you do not do so either. Finally, to help make this session run smoothly, please mute your microphone when you are not speaking.

During this session, we will be generating ideas for a tool that will help people understand their data downloads. Here is an overview of today's activities.

- GDPR/CCPA Overview (2 minutes)
- Exploration of files, then Scavenger Hunt (12-15 minutes)
- Discussion (10-15 minutes)
- Highlight Activity (3-4 minutes)
- Data Visualization 101 (5-7 minutes)
- Sketch activity (10-15 minutes)

But first, let's talk about why you are able to download your data in the first place.

GDPR/CCPA Overview

In response to privacy concerns about online data, two major privacy laws were passed recently. The General Data Protection Regulation came into effect in the European Union in May 2018. GDPR grants users the right to access the data that an online company has about them–a right that you all have exercised as part of this study. Inspired by the GDPR, California produced a similar law, called the California Consumer Protection Act, that went into effect at the beginning of this year. These laws grant other rights, like the right to data portability, the right to erasure of your data, and the right to correct false information about yourself, but today we're going to focus on the right to access. Does anyone have any questions about GDPR or CCPA?

Exploration of Files

So let's talk about your data download. First, did anyone look at their data before this session?

What types of information were you/are you expecting to find in your data download?

What types of information do you want to see?

Please navigate to the next page of the Qualtrics survey.

First, we'll take 1-2 minutes to look at the [index.html, Read Me First pdf file and the "Understanding My Data" link, archive_browser.html, or readme.html] file. This is a visual overview of the folders and files contained in your data download. [This section was omitted for Amazon, which doesn't provide such a file in the order data download.]

Next, take 5 minutes to look through your data on your own. We encourage you to make comments aloud to us and to the other participants as you discover things that you find interesting.

While you look for these items, take time to familiarize yourself with your data, paying particular attention to the information that is included and the format and organization. There are a few guiding questions on Qualtrics. [See Survey 3]

What are your initial reactions?

What surprised you?

What was it like navigating this file?

Did your expectations match the reality of what was contained in the file?

Is there anything you wanted to see but didn't?

Scavenger Hunt

Now we're going to do a short scavenger hunt to help get you acquainted with your data. Please proceed to the next section. You will see a list of items to search for in your data. If you can't find an item, skip it and move on. It's possible that an item may not be in your data at all. While some items can be easily found by looking on the website or app, please only look for the answers in the files that you downloaded. However, you are welcome to use Windows Explorer, Finder, or any other search tool on your computer. You may also use the [index.html, Read Me First pdf file and the "Understanding My Data" link, archive_browser.html, or readme.html] file found in your data download. The goal of this activity is to get you acquainted with your data download. You don't need to write anything down, but you can if you'd like. We'll spend around 5 minutes on this activity.

Again, please comment aloud as you find scavenger hunt items or anything else you find interesting.

Discussion Questions

Next, we have some discussion questions.

Scavenger Hunt

1. How many scavenger hunt items did you find?
2. Did the [index.html, Read Me First pdf file and the "Understanding My Data" link, archive_browser.html, or readme.html] help you with the scavenger hunt?
3. What was it like navigating this file?
4. Was there any information collected about you that you were surprised by? Why?
5. Is there any data you think the company has about you that is missing from these files?

General

1. What are some reasons, if any, you might want to have access to your data?
2. From which websites or apps (social media, online shopping, ride share, etc.) would you be most interested in downloading your data?
3. What pieces or types of data are most important for you to see in a data download?
4. What pieces or types of data are not important for you to see in a data download?
5. How was the process of requesting your data?
6. How did you navigate to the page that gives you access to your data?
7. How long did it take for you to be able to access your data?
8. Were you previously aware of how to navigate through a csv/json/txt file?
9. What records were you looking to find from your data download? Were you able to access them?

Privacy

1. Was there any information collected from you that made you uncomfortable? Why? Do you think this information is useful or important for the company to have?
2. If after seeing this data download, you wanted to share less info with the website, what steps would you take?
3. Do you feel the data about you is accurate?

Design

1. What elements of the data download layout are most intuitive to you, and which were the most difficult to navigate?
2. How was your data separated into folders? Does this organization make sense to you? Can you think of other ways to organize?
3. Was any terminology used in the data download unclear? If so, which terms?
4. Are the file names descriptive?
5. When you think about your interaction with this platform, is it easy to trace your online activity through this data file?

Other

1. How would you feel about adding aggregate statistics to your data–for instance, your average ride cost (Uber), number of 'liked' pages per month (Facebook)
2. How would you feel about a setting that lets you choose different levels of specificity for your report?
3. How would you feel about a tool that helped you make privacy-protective choices based on your data?
4. How would you feel about reminders to do things like delete your data or modify your settings?

Highlight Activity

Please advance to the next page of the survey. For this next activity, we will give you a list of the categories or folder names in your data. Please highlight the ones that would be most important for you to see and understand in your data download. To highlight an item, double click in, then click the word important. You may highlight as many as you'd like. Do not think too hard about your answers. Some categories have confusing or unclear names. Go with your gut.

Data Visualization 101

Please navigate to the next page of the survey. The ultimate goal of our project is to design tools that make it easier for you to understand your data downloads. One way to accomplish this is data visualization in which information is displayed using visuals like charts, graphs, and maps. Let's take a look at a couple of examples of data visualization.

Beautiful News Daily

When raw statistics or numbers are reported in the news, sometimes it can be hard to digest that information. Journalist David McCandless founded a site called information is beautiful dot net, which offers visualizations of the daily news. Take a couple minutes to explore visualizations of the news you find most interesting. As you explore, pay close to attention to examples that synthesize multiple pieces of information to give a more complete or interesting account.

https://informationisbeautiful.net/beautifulnews/

Does anyone want to share a visualization they found particularly interesting or well designed? [If so, ask them to drop the link in the chat.]

Does anyone have an example where multiple types of information were synthesized?

Excel Line Graph

Here is a more basic example of data visualization. Say I wanted to know how many cats I petted each month in 2019. One option would be to look at an excel spreadsheet with this information. However, if I wanted a visual representation of this data, I might take my spreadsheet and convert it to a line chart. I can easily look at this line chart and conclude that March was a great month for cat petting. Data visualization doesn't have to be super complex–it could be a simple graph!

Sketch Activity

Please advance to the next page of the survey. For this last activity, we would like you to imagine that someone designed a tool that generated a visualization of your data. Please sketch your ideal version of this visualization on a piece of paper or using your favorite drawing tool. Do not feel limited to what has been discussed in this session. Don't worry about the quality of your sketch. The goal is to get your ideas across. For example, if you can't draw a unicorn, simply write "picture of unicorn."

You can take several approaches. You could sketch the overall layout of the tool, like the website layout and the different options that the tool provides. Or you could focus on representing a specific type of data, for example, location data. You might also consider how to synthesize multiple pieces of information like we saw in the daily news data example. You are also welcome to take more than one approach. There are a few guiding questions on Qualtrics. [See Survey 3]

Once you're done, please scan in or take a photo of your drawing and upload it to the survey. If you're doing it from your computer, you can click the link. If you're doing it from a phone or other device, you can type in the URL or open your camera to scan the QR code. Type "ready" in the chat when you're done.

Now we're going to share our ideas. Please explain your sketch. If you'd like, we can share your drawing with the group, but you can opt for a verbal explanation only. Who would like to go first?

Closing Remarks

Thank you for participating in our study. You all have been fabulous! Please advance to the final page of the survey to get the completion code, which you will use on Prolific to receive compensation for your participation. If you have any questions about the study, please ask now, or refer to the consent form for contact details. We will stick around for a few minutes.

B Instructions for Downloading Data

Amazon
Part 1: Request Your Data
1. Go to the following URL: https://www.amazon.com/gp/help/customer/display.html?nodeId=G5NBVNN2RHXD5BUW
2. Click the "Request My Data" link.
3. Log in with your Amazon username and password. Then select "Your Orders" from the drop-down menu. Then click "Submit Request."
4. You should see this message: [screenshot of message]
5. Log in to the email associated with your Amazon account. Find the email with the subject line "Your Data Request Confirmation." Click the "Confirm Data Request" button.
6. You should see this message: [screenshot of message]
7. It may take anywhere from a couple hours to a couple days for your data to be ready. Amazon will notify you by email when your data is ready.
Part 2: Download Your Data
8. Login to the email associated with your Amazon account. Find the email from Amazon with the subject line “Your Data Request.” Click the yellow “Download Data” button in the body of the email.
9. You will be redirected to a new page. You may be asked to login to your Amazon account. Click the “Download” button next to all of the files.
10. Make sure you remember where you saved these files. You will need them for part 3 of the study.

Facebook
Part 1: Request Your Data
1. Go to facebook.com.
2. Login with your username and password.
3. Click the blue triangle in the upper right corner.
4. Click "Settings" from the drop down menu.
5. Click "Your Facebook Information" on the left column
6. Click "View" under "Download Your Information."
7. Ensure that "All of my data," "HTML," and "High" are selected. Then click "Create File."
8. You should see this message: [screenshot of message]
9. It may take anywhere from a couple hours to a couple days for your data to be ready.
Part 2: Download Your Data
10. Facebook will notify you when your data is ready either by email or via a Facebook notification.
    - **Option 1**: Login to the email associated with your Facebook account. Find the email with the subject line “Your Facebook information file is ready.” Click the “Download Your Information” link found in the body of the email.
    - **Option 2:** Click on the Facebook notification that looks like this: [screenshot of notification]
11. You will be redirected to a new page. You may be asked to login to your Facebook account. Click “Download” on your most recent file.
12. Make sure you remember where you saved this folder. You will need it for part 3 of the study.

Google
Part 1: Request Your Data
1. Go to https://myaccount.google.com/?utm_source=sign_in_no_continue
2. Log in with your username and password. Please use your primary Google account. Note: if you are already signed in, you can skip this step.
3. Click “Data & Personalization” from the left column
4. Scroll down until you see “Download, delete, or make a plan for your data.” Click “Download your data.”
5. Click “Deselect all.”
6. Scroll down until you see “Location History.” Check the box.
7. Scroll down until you see “My Activity.” Check the box.
8. Click the “All activity data included” button.
9. Click the “Deselect All” button.
10. Check the “Search” box.
11. Press the “OK” button.
12. Click “Next Step” at the bottom right corner.
13. Leave all the presets alone. The page should look like this: [screenshot of page]
14. Click “Create export.”
15. It may take anywhere from a couple hours to a couple days for your data to be ready. Google will notify you by email when your data is ready.
16. Note: Some people have reported that they didn’t receive an email. If you haven’t received an email after a couple days, go to https://takeout.google.com to see if your data is ready.
Part 2: Download Your Data
17. You will receive a link to the email address associated with your account. Follow this link and press “Download.”
18. Make sure you remember where you saved this folder. You will need it for part 3 of the study.

Spotify
Part 1: Request Your Data
1. Go to spotify.com.
2. Log in with your username and password.
3. Click "Account" under the "Profile" menu in the top right corner.
4. Click "Privacy settings" from the left column.
5. Scroll down to the "Download your data" section. Click the "Request" button.
6. Log in to the email associated with your Spotify account. Find the email with the subject line "Confirm your Spotify data request." Click the "confirm" button.

7. You should see this message: [screenshot of message]
8. It may take several days for your data to be ready. Spotify will notify you by email when your data is ready.
Part 2: Download Your Data
9. Login to the email associated with your Spotify account. Find the email with the subject line " Your Spotify personal data is ready to download." Click the green "Download " button in the body of the email.
10. Type in the password to your Spotify account and click “verify.” The download will start automatically.
11. Make sure you remember where you saved this file. You will need it for part 3 of the study.

Uber
Part 1: Request Your Data
1. Go to the following URL: https://auth.uber.com/login/?breeze_local_zone=dca1&next_url=https%3A%2F%2Fmyprivacy.uber.com%2Fprivacy%2Fexploreyourdata%2Fdownload%3F_ga%3D2.160201528.441384756.1587066962-1367774538.1587066962&state=K5fXVafN4vyOBujPSOoPLCsftsZFkaPRRmI81J_NvwY%3D
2. Enter your email address.
3. Enter your password.
4. Enter your phone number, and then the 4-digit code.
5. Click "Request Your Data."
6. You should see this message: [screenshot of message]
7. It may take several days for your data to be ready. Uber will notify you by email when your data is ready.
Part 2: Download Your Data
8. Login to the email associated with your Uber account. Find the email with the subject line “Your Uber data is ready for download.” Click the green “Go to Download Page” button in the body of the email.
9. You will be redirected to a new page. You may be asked to login to your Uber account. Click the blue “Download” button.
10. Make sure you remember where you saved this folder. You will need it for part 3 of the study.

YouTube
Part 1: Request Your Data
1. Go to https://myaccount.google.com/?utm_source=sign_in_no_continue
2. Sign in with your username and password. Please use the primary Google account you use to access YouTube.
3. Click "Data & Personalization" from the left column.
4. Scroll down until you see "Download, delete, or make a plan for your data." Click "Download your data."
5. Click "Deselect all."
6. Scroll down until you see "YouTube and YouTube Music." Check the box.
7. Click "Next Step" at the bottom right corner.
8. Leave all the presets alone. The page should look like this: [screenshot of page]
9. Click "Create export"
10. It may take anywhere from a couple hours to a couple days for your data to be ready. Google will notify you by email when your data is ready.
11. Note: Some people have reported that they didn’t receive an email. If you haven’t received an email after a couple days, go to https://takeout.google.com to see if your data is ready.
Part 2: Download Your Data
12. You will receive a link to the email address associated with your account. Follow this link and press “Download.”
13. Make sure you remember where you saved this folder. You will need it for part 3 of the study.

C Contents of Data Downloads

Category	Amazon	Facebook	Google	Spotify	Uber	YouTube
Communications	Gift Messages	Comments, Sent friend requests, Posts, Like and reactions, Messages, Pokes, Stories	–	–	Info on support conversations with Uber	Comments
Inferences	–	Off-Facebook activity, Ad interests, Advertisers interacted with, Information submitted to advertisers, Advertisers with a contact list of your info	–	List of market segments user is associated with	–	–
Locations	Billing address, Shipping address	Primary location, Profile current city, IP where you've logged in, IP addresses of user device for login, your places	Location, Latitude, Longitude, timestamp, velocity, altitude, activity at location, type of activity	User address, Payment country and card postal code, Family plan address, Car thing shipping address	Locations and times at which a trip (either using Uber Rider or Uber Jump) was started and ended	–
Payment Data	Payment instrument type for orders and subscriptions	Facebook Pay payment history and payment methods	–	Details of payment data	Payment method info	–
Primary Usage Data	Orders and Subscriptions info	Events, Posts, Stories, Following, Friends, Groups,Like and reactions, Marketplace activity, Pokes, polls voted on, support correspondence, Photos and videos uploads, search history, account activity	Searches	Follow-ing/Followers data, search queries, streaming history, playlists	Uber rider trips history, Uber jump bike rides history, Uber eats order history	Likes, playlists, video uploads, subscriptions
Search History	–	Time-stamped searches	Time-stamped searches	List of searches with date and time, type of device/ platform used to make search	–	Timestamped searches
User Profile	–	Name, Previous names, Emails, Birthday, Gender, Current City, Hometown, Education, Work experiences, Phone numbers, Bio, Registration timestamp, Profile update history	–	Username, email address, address, mobile number, mobile operator, mobile brand, gender, birthday, registration date, Facebook user ID	Name, email address, mobile number, ratings, and registration date	–
Voice Data	–	Voice recording and transcript	–	List of voice input commands	–	–

Table 2: Our informal categorization of data contained in Amazon, Facebook, Google, Spotify, Uber, and YouTube data downloads.

Facial Recognition: Understanding Privacy Concerns and Attitudes Across Increasingly Diverse Deployment Scenarios

Shikun Zhang
Carnegie Mellon University
Pittsburgh, PA, USA
shikunz@andrew.cmu.edu

Yuanyuan Feng
Carnegie Mellon University
Pittsburgh, PA, USA
yuanyuanfeng@cmu.edu

Norman Sadeh
Carnegie Mellon University
Pittsburgh, PA, USA
sadeh@cs.cmu.edu

Abstract

The rapid growth of facial recognition technology across ever more diverse contexts calls for a better understanding of how people feel about these deployments — whether they see value in them or are concerned about their privacy, and to what extent they have generally grown accustomed to them. We present a qualitative analysis of data gathered as part of a 10-day experience sampling study with 123 participants who were presented with realistic deployment scenarios of facial recognition as they went about their daily lives. Responses capturing their attitudes towards these deployments were collected both in situ and through daily evening surveys, in which participants were asked to reflect on their experiences and reactions. Ten follow-up interviews were conducted to further triangulate the data from the study. Our results highlight both the perceived benefits and concerns people express when faced with different facial recognition deployment scenarios. Participants reported concerns about the accuracy of the technology, including possible bias in its analysis, privacy concerns about the type of information being collected or inferred, and more generally, the dragnet effect resulting from the widespread deployment. Based on our findings, we discuss strategies and guidelines for informing the deployment of facial recognition, particularly focusing on ensuring that people are given adequate levels of transparency and control.

1 Introduction

We live in a world full of cameras, from traditional closed-circuit televisions to the latest motion-sensing wireless IP cameras. According to a report by IHS Markit, a total of over one billion cameras are expected to be deployed worldwide by 2021 [25]. Existing security and surveillance cameras can be easily augmented with facial recognition, a type of artificial intelligence (AI)-enabled video analytics technology that has become increasingly accurate with recent advances in deep learning and computer vision [43]. The U.S. Government Accountability Office (GAO) broadly defines facial recognition technology as computer applications that (1) detect faces in an image or video, (2) estimate a person's demographic characteristics (e.g., age, race, gender) (3) verify a person's identity by accepting or denying the claimed identity, and (4) identify an individual by matching an image of them to a database of known people [105]. Extensions of facial recognition also include facial expression recognition [87], mood detection, scene detection (e.g., identifying petty crime [85]), and more. In this paper, we adopt this broader definition of facial recognition.

In recent years, facial recognition has been widely deployed in public places, such as airports for security and surveillance purposes [42, 103], department stores for automatic detection of known shoplifters, rental car companies for self-checkout [37, 74]. While facial recognition technology can contribute to security, productivity, convenience, and more, its broad deployment also gives rise to serious privacy concerns [97]. These concerns have prompted increased scrutiny from both privacy advocates and regulators [24, 30, 60]. Recent studies have also reported limitations and flaws of facial recognition technology, including unsatisfactory levels of accuracy as well as bias towards underrepresented demographic groups and members of the LGBTQ+ community [5, 47, 59, 81]. Both policymakers and researchers have also expressed concerns about abusive uses of the technology, e.g., non-consensual surveillance [48, 49].

Our research focuses on the perceptions and attitudes of people (or "data subjects") whose presence and activities can be captured by facial recognition technologies. This paper describes the results of an exploratory qualitative analysis of responses gathered as part of a 10-day experience sam-

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

pling study. The study involved asking participants to install a study app on their regular smartphones and using the app to present them with a range of realistic facial recognition scenarios at venues they visited during their everyday activities. The app was used to collect their reactions to these different scenarios. Data collected in situ was supplemented with additional information collected as part of a daily evening survey, in which participants were asked to review each of the deployment scenarios presented to them during the day and answer additional questions. Moreover, we analyzed 123 participants' post-survey responses and also interviewed 10 of them. This paper is the sequel to another publication on this study, where we presented a quantitative analysis of participants' privacy preferences and expectations in responses to these scenarios [115].[1] Through an in-depth analysis of the qualitative data collected from this study, we aim to develop a more holistic understanding of people's perception of the benefits and concerns associated with the diverse deployment scenarios considered in this study. In particular, we further contextualize participants' perception of privacy risks associated with facial recognition and explore their concerns about the limitations and bias found in some of these systems.

This article's contributions fall under three broad categories:

- We present an in-depth qualitative analysis of lay people's perceptions towards facial recognition. Our qualitative dataset contains both interview data and 123 participants' free-text responses over a 10-day period, which provides a comprehensive view of individuals' perceptions towards facial recognition.
- To our knowledge, this is the first qualitative study that uses carefully designed and realistic facial recognition deployment scenarios and differentiates between diverse attributes of the technologies (e.g., purpose, venue, type of analysis, data sharing) and to do so *in situ*, as participants went about their regular everyday activities.
- Based on our results, we propose guidelines and design recommendations for trustworthy deployments of facial recognition technology.

2 Related Work

2.1 Facial Recognition and Algorithmic Bias

Facial Recognition (FR) and its wide range of applications have been a prevailing research topic for decades. Traditional FR methods are mostly feature-based and are limited in their discriminant power [6, 14, 50, 67, 113]. Recent deep learning approaches have significantly boosted the performance of facial recognition models [26, 76, 90, 104], enabling it to approach and surpass human performance on FR benchmarks [45, 55, 58]. Despite the impressive progress, there are still many problems with FR. A series of reports on testing commercial facial recognition software conducted by the National Institute of Standards and Technology (NIST) revealed that software accuracy variations and potential bias existed for different demographic groups [4, 43, 44]. Several studies have tried to quantify the demographic biases of some of these deep learning models [17, 56, 81], with sources of bias attributed to unrepresentative data distributions in training sets [56] and to the use of certain optimization methods [93]. Besides issues related to accuracy and bias, prior studies have also questioned the effectiveness of emotion detection, which falls under the broad definition of facial recognition, exposing problems with classifiers trained on artificial displays of emotions failing to capture people's true inner states [12, 69]. These limitations can in turn lead to the mistreatment of certain demographic groups, exposing them to higher individual or societal risks, or impeding their access to some services [111]. Even though these problems have been acknowledged by the computing community [5] and legal scholars [53, 112], no research has been conducted to understand people's awareness and perception of these limitations and the risks they entail. Our work aims to fill this gap.

2.2 Attitudes towards Facial Recognition

A few prior studies have examined people's attitudes towards facial recognition through surveys [21, 97, 100, 101]. The Pew Research Center conducted a nationally representative survey on Americans' awareness and acceptance of facial recognition. They found that Americans in general trust law enforcement to use facial recognition responsibly more than technology companies and advertisers and that these attitudes also vary across demographic groups [97]. Another study further analyzed the Pew survey data and focused on gendered perceptions of workplace surveillance. This study found that women were less likely to accept the use of facial recognition in the workplace [100]. The Center for Data Innovation also conducted a national online poll through Google Surveys and found that fewer Americans think the government should limit the use of facial recognition [21]. A few interview studies have focused on specific functionalities of facial recognition [8] and the impact of facial recognition technology on marginalized demographic groups [47]. Hamidi et al. found transgender individuals have overwhelmingly negative attitudes towards recognition algorithms that automatically detect gender [47]. Andalibi et al. discussed users' attitudes towards emotion recognition technology, including perceptions of individual and societal risks [8]. Our work, which does not focus on the relationship between demographics and attitudes, sheds light on people's concerns about facial recognition across a variety of scenarios without targeting

[1] See also [116] for results exploring the use of machine learning models to help predict people's privacy preferences (i.e., opt-in/opt-out preferences for different scenarios) and alleviate the user burden of exercising privacy choices.

any particular demographic group. Facial recognition has also attracted the attention of law scholars who have closely examined the legal and ethical issues of the emerging facial recognition through a slew of law review articles [53, 72, 112]. Our work complements these legal reviews by presenting and analyzing data collected from our study participants, relying on their own accounts of perceived threats and benefits associated with these deployments in realistic contexts experienced as part of their regular everyday activities.

2.3 Privacy Challenges of Facial Recognition

Facial recognition technology can be used to capture a variety of sensitive information about people, from biometric data (e.g., facial features and body pose) [38, 90, 104] to information about people's activities (e.g., where they are, whom they are with, and what they do) [38, 114] all the way to their emotions (e.g., attentive, depressed, and surprised) [64]. While people may notice some cameras, they have no way of knowing how captured footage is being processed (e.g., what types of algorithms might be run and for what purpose) and what happens to the data being captured (e.g., whom it is shared with and for how long data might be retained). The loss of privacy resulting from the deployment of this technology has been a common thread in the literature [19, 70, 78, 79]. Researchers have examined technical solutions to safeguard user data [31, 32, 34, 78, 89], including algorithms to avoid being tracked by facial recognition [94, 95], and systems to enable real-time opt-out of facial recognition systems [27, 28, 88]. But how to increase transparency around data privacy remains an unsolved issue [22, 82, 83].

In this paper, we explore three research questions:

- RQ1: What are users' attitudes towards facial recognition technology, and why?
- RQ2: What are some benefits and concerns people associate with facial recognition deployment scenarios?
- RQ3: What recommendations can we develop for the trustworthy deployment of facial recognition?

3 Methodology

3.1 Study Design

Prior work shows that context plays a critical role in influencing people's privacy attitudes and decisions [75]. In order to solicit realistic participant feedback, we designed an experience sampling study to collect people's responses to a variety of facial recognition deployments (or "scenarios") in the context of their regular everyday activities. The experience sampling method [51] has been successfully used in many real-life studies [20, 39, 54, 61, 84, 106, 107], enhancing the ecological validity of the results [13, 92].

In the 10-day experience sampling study, we presented participants with facial recognition scenarios that were likely to happen at places they visited as part of their daily activities. For example, when a participant visited a gym, they may be presented with a scenario where facial recognition was used to track their attendance. The scenarios included in the study were informed by an extensive survey of news articles about real-world deployments of facial recognition in a variety of contexts, i.e., identification of known criminals [2, 23, 40, 57], petty crime detection [85], operation optimization by businesses [71, 77, 86], demographic-based advertising [9, 35, 98], advertising based on reactions [15, 18, 91], engagement detection [63, 68, 110], ID/loyalty card replacement [10, 33, 73, 96], attendance tracking [3, 11, 41], health-related predictions [7, 66, 80], productivity predictions [29, 62], and medical diagnoses [1, 36, 46, 65].

3.2 Study Procedures

The 10-day study was carried out in the following steps. First, eligible participants who completed the consent forms could download the in-house study app from the Google Play Store. Second, while participants went about their regular daily activities, the study app collected the GPS location of their smartphones. As participants visited places for which we had plausible scenarios, the app would send them a push notification, prompting them to complete a short survey on a facial recognition scenario pertaining to their location. Third, at the end of each day, participants also received an email in the evening to answer a daily summary web survey ("evening review"). This web survey showed participants the places they visited when they received notifications, probed reasons for their in-situ answers, and asked additional questions. See Appendix 7.4 for screenshots of the app and an example of the evening review. Fourth, after completing 10 days of evening reviews, participants answered a post-survey where they provided open-ended text responses about their attitudes on facial recognition technology and their perceived beneficial and concerning contexts where facial recognition was applied. Fifth, we conducted semi-structured interviews with 10 participants over online video conferencing software (e.g., Skype, Google Hangouts) after they have completed the study. The full text of the post-survey, the scenarios presented during the study, and the interview scripts can be found in the Appendix.

3.3 Recruitment and Participants

We recruited participants from both online and offline channels. Our recruitment messages were posted on a variety of online platforms, including local online forums (i.e., Craigslist and Reddit), a university-based research platform, and a promotional Facebook advertisement. We also put up flyers on bus stops and local community bulletin boards. A short screening survey was used to determine participants' eligibility

(aged 18 or older, able to speak English, using an Android smartphone with a data plan). We also collected demographic information such as age, gender, and occupation in the screening survey. We avoided convenience samples of university students and collected data from a diverse pool of participants. A total of 164 participants downloaded our study app, and 123 of them completed our 10-day study and the post-survey. The demographics of the 123 participants is shown in Table 1 and 2. We sent out 17 invitations to participants who showed interest in participating in the follow-up interview and conducted online interviews with 10 participants who responded. This study was approved by our university's IRB and the human research protection office of the funding agency.

3.4 Qualitative Dataset

In this work, we focused on analyzing the qualitative dataset collected from the 10-day experience sampling study. The dataset includes 2,562 entries of text responses from participants' daily summaries, 1,230 entries of text responses in the post-survey, and 10 interview transcripts. In order to answer the research questions, it is crucial that the qualitative data collected reflects participants' attitudes towards facial recognition. Since we adopted an experience sampling method presenting realistic scenarios of facial recognition to participants over 10 days, we believe the data collected following these contextual cues would capture participants' perceptions and attitudes. We did not report other quantitative data collected from the experience sampling study since they are not the focus of this paper.

3.5 Interview Data Analysis

The interviews ranged from 26 to 40 minutes (mean=33) and were fully transcribed. A total of 326 minutes of transcripts were analyzed. One author first read and familiarized herself with all the transcripts. She then applied thematic analysis [16] to open code the transcripts. The second author met with the first author regularly to iterate on the themes.

3.6 Content Analysis of Textual Responses

From the 10-day study, we collected 2,562 entries of text responses from participants' daily summaries and 1,230 entries from the post-survey. In the post-survey, there were 10 open-ended questions. The first question was "What is the first thing that comes to your mind when you think about facial recognition technology?" We coded the sentiment (i.e., positive, negative, neutral, mixed) in each response.

We included two questions in the post-survey asking participants' perceived beneficial and concerning contexts to use facial recognition technology. We also asked questions eliciting participants' privacy concerns about facial recognition deployment scenarios. After reading the survey responses, we realized many participants shared their attitudes and experiences with facial recognition deployment scenarios regardless of to which question they were responding. Since the daily summaries were also addressing similar issues, in our analysis, we broke down the boundaries between the data sources and conducted a content analysis [102] of all the participants' 3792 textual responses.

Two authors started from inductive coding [16] to extract codes that show participants' perceived benefits or concerns about facial recognition technology and developed a codebook. In total, we summarized 13 main codes with 32 subcodes about the benefits of facial recognition and 19 main codes with 40 subcodes about the concerns. In the end, we used a deductive coding approach, applying the codebook to the entire dataset. Two authors independently coded all data and met to resolve any discrepancies.

4 Findings

In this section, we present findings from qualitative analysis of interview and textual response data collected from evening reviews of in-situ scenarios participants received. We first present findings on participants' attitudes towards facial recognition technology and the reasons behind their attitudes. We then show the perceived beneficial and concerning contexts of facial recognition usage. We also unveil participants' concerns about the use of facial recognition, with a particular focus on privacy-specific concerns, as they are among the most prominent themes. Finally, we flesh out participants' proposed actions in responses to these deployment scenarios.

4.1 Impressions of Facial Recognition

We first present findings on participants' sentiment towards facial recognition technology. This is based on our coding of sentiment in participants' responses to the first question in the post-survey: "What is the first thing that comes to your mind when you think about facial recognition technology?"

4.1.1 Participants tend to be more negative towards FR

We observed that participants tended to be more negative towards the use of facial recognition: 51 (42%) participants displayed negative impressions while only 13 (11%) expressed positive sentiments. The negative connotation mostly revolves around problems of the technology, like the infringement on their right to privacy. Those negative first impressions also echo entrenched perceptions on problematic usages and privacy risks of facial recognition that are revealed in our subsequent analysis in Section 4.4.2 and 4.5.

Among the 13 participants with positive impressions, most praised facial recognition's usefulness, like its ability to increase public safety and catch criminals. A few also mentioned the *"advancement in technology"* (P36, positive). We

Gender	%	Age	%	Education	%	Income	%	Marital Status	%
Female	57.7	18-24 years old	8.1	Some high school	.8	Less than $25,000	14.6	Single, never married	50.4
Male	40.7	25-34 years old	54.5	High School	4.1	$25,000 to $34,999	14.6	Married	41.5
Other	1.6	35-44 years old	23.6	Some college	13.8	$35,000 to $49,999	9.8	Separated	1.6
		45-54 years old	8.1	Associate's degree	7.3	$50,000 to $74,999	22.0	Divorced	3.3
		55-64 years old	3.3	Bachelor's Degree	35.0	$75,000 to $99,999	14.6	Widowed	0.8
		65-74 years old	2.4	Master's Degree	23.6	$100,000 to $149,999	14.6	I prefer not to answer	2.4
				More than Master's Degree	12.8	$150,000 to $249,999	2.4		
				Other	1.6	I prefer not to answer	7.3		

Table 1: Survey participant demographics and respective %

Occupation	%	Occupation	%
Business, or sales	12.2	Legal	3.3
Administrative support	9.8	Other	3.3
Scientist	8.9	Graduate student	2.4
Service	8.1	Skilled labor	2.4
Education	8.1	Homemaker	2.4
Computer engineer or IT	7.3	Retired	2.4
Other salaried contractor	7.3	Government	1.6
Engineer in other fields	6.5	Prefer not to say	1.6
Medical	6.5	Art or writing	.8
Unemployed	4.1	College student	.8

Table 2: Occupations of survey participants and respective %

also noted a mixed perspective of facial recognition from 11 (9%) respondents: *"It's invasive and big brother esque. It can provide good information for law enforcement but is easily abusable"* (P83, mixed). 48 participants (39%) indicated their neutral impressions typically by describing main use cases or depicting how facial recognition works: *"the ability of computers to see normal people in plain view and identify their identity. This can then be passed to another decision-making system for a distinct purpose: law enforcement, advertising, efficiency, etc."* (P12, neutral).

4.1.2 Participant views may be influenced by media portrayals

A few concepts also emerged from these responses, mostly related to media portrayals of facial recognition. Some participants were reminded of what they have watched in the movies or crime shows relating to facial recognition: *"I think of face scanners and searches people do when looking for criminals in crime tv shows and movies"* (P42, neutral). Other respondents made references to a dystopian world, with many citing the concept of Big Brother from the book 1984 — *"Cyberpunk dystopias, "Big Brother," and similar instances in fiction, satire, and socio-political discussion about invasion of privacy on the part of powerful political and economic entities"* (P39, negative). China was brought up 7 times as the example of a surveillance state, which was associated with more negative sentiments (5 out of 7) than neutral tones (2 out of 7). For example, P80 alluded to a negative use case, *"China and the way they micromanage their citizens lives,"* and P5 expressed a more neutral impression: *"I think of China because the only times I've seen it on the news, it was being used in China."*

In summary, respondents expressed more negative views about facial recognition than positive ones. Many were wary about potential problems linked to the technology. Around a quarter of participants' views were influenced by the media portrayal of facial recognition (e.g., news, movies, and books).

4.2 Beneficial and Concerning Contexts

We present findings on users' perceived beneficial and concerning use of facial recognition. This is based on the deductive coding of textual responses to the questions asking participants to identify up to 5 contexts each where they found the use of facial recognition technology to be beneficial and concerning. On average, each participant identified 2.7 ± 1.4 beneficial contexts, and 3.0 ± 1.4 concerning contexts. A paired Wilcoxon signed-rank test showed that participants recorded significantly more concerning contexts than beneficial contexts ($Z = 2.65, p < 0.01, r = 0.24$).

The findings are organized based on the major codes in the codebook, as shown in Table 3. These codes were further categorized into two groups: purposes for using facial recognition and entities that use facial recognition. We first report beneficial and concerning purposes in this subsection.

4.2.1 Beneficial purposes: security, authentication, and commerce

The majority (104 out of 123) of participants reported that security is a beneficial context for facial recognition. Among those, 42% thought that facial recognition could increase public security in general, and 32% thought that it is beneficial to use facial recognition to identify and catch criminals. Another important context for security, raised by 20%, is to find missing individuals. For example, P26 mentioned that facial recognition could be helpful in *"locating missing/abducted children and adults."* 13% of them also mentioned that facial

recognition could be beneficial to deter crime, as expressed by P27 *"in public, especially in isolated places like parking garages, to help preserve women's safety."* Another context for facial recognition that 51 participants (42%) identified as beneficial is authentication. About half of them (24 out of 51) stated that facial recognition could be used to replace IDs and confirm identity. 31% mentioned that it could be used to log in devices and/or replace passwords. A quarter maintained that facial recognition could be useful to grant access in secured locations, which P46 described as *"helping identify people in high-security areas."* 14% considered authentication in stores via facial recognition as a way to replace membership or reward cards to be beneficial as well. A sizable minority (27 out of 123 — 22%) of participants also saw merits in leveraging facial recognition in commercial settings; using facial recognition to improve services and tailor customer experiences was deemed beneficial by about half of those 27 participants, for example, in contexts like *"relocating people between the crowded check-out areas"* (P63) and *"customization of service based on who you are and known preferences"* (P55). Others considered marketing and tailored advertisement of potential benefit, like in *"retail scenarios (catered advertising)"* (P46) and *"providing information to retail companies about their customers"*(P111).

4.2.2 Concerning purposes: advertisement, profiling, and prediction

Most participants (64%) raised concerns about various purposes for which facial recognition is used. Specifically, 36 out of 123 (29%) participants found using facial recognition for advertisement troubling: P117 said, *"It can be used for marketing and branding purposes that are generally antagonistic."* 18 participants were concerned about facial recognition used for profiling — *"using it to profile someone based on race or gender"* (P21). 17 respondents found *"when emotion recognition is in use"* to be concerning. 12 participants (10%) were specifically against their data being sold for profit *"random companies selling and profiting off of it"* (P40). 11 were worried about use cases of facial recognition that involves predicting or estimating intentions or behaviors — *"Any assessments that are psychologically based since there is a lot that could be wrongly inferred by only taking into account visual data"* (P12).

4.3 Beneficial and Concerning Entities

The right-hand side of Table 3 shows the percentages of participants who identified different entities (law/government, employers, etc.) as beneficial and/or concerning when they deploy facial recognition.

Purpose				Entity			
Beneficial		Concerning		Beneficial		Concerning	
Code	%	Code	%	Code	%	Code	%
Security	84.6	Ads	29.3	Law/Gov	14.6	Law/Gov	18.7
Authentication	41.5	Profiling	14.6	Public	11.4	Employer	17.1
Commercial	22.0	Emotion	13.8	Health	8.1	Business	15.4
Personal	9.8	Profit	9.6	Employer	5.7	Insurer	14.6
Other	8.1	Predictive	8.9	Myself	5.7	Health	7.3
		Security	5.7	Business	4.9		

Table 3: Codes from Content Analysis and the Percentages of Participants Who Mentioned Them

4.3.1 Weighing between beneficial versus concerning

It is interesting to observe that law enforcement/the government were deemed concerning and beneficial both by a sizable number of respondents, which is also similar in the case of health-related entities (e.g., hospitals and clinics). The neck-and-neck numbers seem to suggest that those entities entail both rather apparent pros and cons of using facial recognition. For example, *"law enforcement falsely accusing someone"* (P83) is rather concerning, while facial recognition aids *"law enforcement to track and apprehend criminals"* (P42) is clearly beneficial. On the other hand, significantly more participants considered businesses, employers, or health insurers' use of facial recognition more concerning than beneficial. More participants see harm than benefit brought by facial recognition usages by these entities, as elaborated by P59, *"The data collected seems worth more to the company than any coupons could possibly be for me."*

4.3.2 Attributes influencing attitudes towards entities

The interview data revealed in-depth deliberations participants had while weighing various entities obtaining their facial recognition information. **Trust** was one of the factors that can erase participants' doubts about potentially questionable facial recognition usages. Two interviewees explained why they trust their employer or the government/law enforcement, therefore trusting their use of facial recognition. P55 explained, *"I trust my manager personally to have my own interests in heart...Right now, personally, I have a good relationship with my manager and with the company. So I am pretty comfortable with what they do, decide to do, and feel like that they are not going to use it against me."* Believing in the democratic government, P57 maintained, *"The government supposedly is "by the people, for the people" as supposed to private corporations...So if it's used by law enforcement, I am a bit more comfortable with that."*

More evidence on trust being an influential factor also emerged in the answers from evening surveys: *"Because law enforcement and the government have a history of using data for purposes other than what they were intended for or what we were told it was for"*(P26), *"I don't trust insurance companies to make fair decisions"*(P116), *"I trust the library mostly not to do anything bad with the video"* (P97), *"This is*

a large entity that I trust"(P51), and etc.

Besides trust, whether entities that deploy facial recognition have **control** over data subjects is another important attribute. Three interviewees were reserved about their employer or the government using this technology as those entities intrinsically have more **control** over them. In their views, facial recognition can be used against them by powerful entities, such as governments, employers, and big corporations, as expressed in the following quotes.

"I am used to people that advertise to me, trying to sell me something...I have more control over that relationship because I can always turn down buying something, even with coercive tactics that are manipulative. But with my boss or the government, I don't have the power in that relationship at all. So it's more information for them that they can use against me basically." — P50

"I mean whoever's behind it [facial recognition] has more data and information, what people need, what individual person wants, and how to best serve the people around, like get their product to the people. And also they have more control...over their customers." — P52

Three interviewees were worried about advertisers' or corporations' usage that could decrease their sense of **autonomy**. Thanks to facial recognition technologies, businesses would leverage highly fine-grained and even real-time data to improve their marketing techniques. For example, P56 expressed her concern, *"With the ability to read your reactions and then be able to market responses specifically to you, you might be losing some free choice. Because they are able to pinpoint and push harder things they think are important to you, because you are reacting to them, they can get real-time reactions to products...They can start using terms that look like something and trick you into buying something."* Such practices can be manipulative and encroach on people's freedom.

4.4 Concerns About Facial Recognition

4.4.1 Participants were concerned about facial recognition even for anonymous demographic detection

Current facial recognition software enables different levels of identification: some can recognize the shape of faces and humans; some can detect specific demographic features; others can match faces to images of people stored in databases. Demographic detection has been used in contexts like targeted advertising and marketing [9, 35, 40, 98].

When designing the study, we initially conjectured that people would be more comfortable with anonymous demographic detection than personally identifiable detection. Nonetheless, 9% of participants expressed reservations about using anonymous demographic detection for advertising as they saw it as a form of profiling. P50 explicitly pointed out, *"I was also pretty concerned when the notifications popped up about predicting purchases based on racial classifications because that just seemed very racist to me. Just because someone is African American or Hispanic, you can't predict what they are going to want to buy based on their race; that seems a really not very good policy."*

Others were really against gender-based advertising. For example, P50 mentioned, *"And gender, there is such a spectrum, just because you're female, that doesn't mean you are going to wanna wear makeup or buy pretty dresses. Same thing for guys. I just think lumping every person into a classification is over-generalized; you are going to miss people."* Some participants questioned the efficacy of advertising based on gender and race, *" I wouldn't think it will be very accurate, you could target something to me being white that would not at all relate to me still based on that one factor. But it may relate to a non-white person. I think it wouldn't even be accurate. I think you need a lot more than race and gender to advertise to someone effectively*" (P106). This type of practices, even though beneficial at times, can also reinforce existing gender and cultural stereotypes — *"I understand that some ethnic groups might benefit from this (for instance, African American women need specific hair care products that aren't always easy to find.) But I am concerned about the potential for misuse of this technology to discriminate. Also, people don't always "look like" the racial or ethnic background with which they identify"* (P27).

Some participants, including some parents, were leery of age-based advertising, especially worrying about kids being susceptible to those practices. *"Things are marketed to kids nowadays, and kids can buy things on apps without their parents even knowing...I don't think they should be marketed towards kids necessarily"* (P50). We also observed reservations from participants who were afraid of being labeled as a specific demographic group, such as religious groups. P53 said, *"I think it is kind of dangerous to pinpoint one person as part of a group vs. just the individual. So I think the times I was most concerned during the research was when I would go to someplace that was religious[ly] affiliated or like a nonprofit organization. If there was a video of me and my friends maybe at a church or at a Jewish organization. Does that put us more in danger if we are associated with that group? I feel like there is this danger of having a label placed on you, and if the wrong person gets that information, and that could be a catalyst for violence."* P89 summarized her feelings towards demographic-based facial recognition, *"I do think it will divide us more if they are targeting specifically based on what you look like, not even necessarily your profile and who you are...I think it just gives an overall weird and gross feeling, especially in today's society where it comes up a lot."*

4.4.2 Participants were worried about incorrect detection and interpretation

About a third of the participants reported their concerns about the accuracy of facial recognition during the study. Some

were worried about the technology not accurate enough and could make *"mistakes in the face recognition (twins, relatives)"* (P65). One interviewee P107 shared his firsthand experience with inaccurate facial recognition in details, *"I don't know how accurate they would be based on stuff that I have tested out before. Like even with having a beard, it throws off a lot of things that try to guess things. Actually, at work, just for fun, one of the guys had it. It is for visually impaired people who are blind. It scans anything and tells you what it is. It scans faces and got a lot of people like "39 male," and it would be really close, but when it comes to me, it would say 40 where I am 25. It would say frowning even though I am smiling because of it tracking the mustache...if they are trying to pick up people with negative emotions for security purposes, maybe it could be pretty wrong."* Others also echoed their doubts about the accuracy of emotion detection, like P68 *"I don't see how it (emotion analysis) could be that accurate unless you are monitoring what I am saying too. Like I said, I went through a breakup that week, and sometimes I was not in a good mood no matter where I was, no matter how good the food was. How are they supposed to know? It just seemed like it was an unnecessary addition that wouldn't end up being very accurate."*

In addition to questioning how accurate facial recognition can be, some participants also argued that seemingly suspicious behavior, when viewed out of context, can be misinterpreted by those systems, potentially resulting in grave consequences. For example, P53 described one such scenario in her friend's life that could be misconstrued, *" I think a lot of the times like my friend she locked herself out of her apartment this past weekend, so she tried to jump in through her window. So if a recognition program saw that, they might think that it is a thief or criminal or whatever. And that is not the case. She is not breaking into her own house. It needs to be able to interpret scenarios correctly. It needs to be able to have a context for them. Not just to assume that something looks like a criminal act is a criminal act."* Similarly, P68 gave another example, *"I think it could misinterpret scenarios, it could misinterpret the guy trying to break into his own car to get his keys out, or the boyfriend putting his hand in the girlfriend's pocket."* An interviewee P57 was worried about such inaccuracies leading to deadly consequences — *"because if someone was marked for shoplifting and they didn't do, that could cost a lot of trouble, in some scenarios that could cost someone's life."*

4.4.3 Participants were concerned about racial and other biases introduced by facial recognition

One-tenth of our participants reported being concerned about potential bias in the facial recognition systems, especially about the deep implications it might have on minority groups. Many were worried that racial bias in these algorithms could exacerbate the entrenched bias and infringe upon the rights of those impacted groups. Two interviewees' elaborated accounts provide us with more insights: P68 stated, *"Any system I've seen has inevitably been used only to profile people of color and the LGBTQ+ community. I think even if we have this surveillance, somebody is like, "Oh, it is just gonna automatically detect petty crimes." The reality is that it will still be looking harder at a black person and their actions to see if that is a petty crime than it could with a white person. I still think at the end of the day, a human is gonna analyze the data. I think you still have a lot of misidentification where people of color and LGBTQ+ community members are going to be scrutinized more strongly, not given the benefit of the doubt that white people are."* Similarly, P53 noted, *" I wouldn't want a program like that to decide that for example, a black man equals thief or even to give a warning sign to a program to flag that because that is not the case. So I think that is the danger of having that type of use for facial recognition. I think it can too easily be biased, intentionally or unintentionally. The person programming it might think that they might have statistics to back up the demographics of thieves or demographics of criminals, but I don't think that is a good way of deciding who is or who is not a criminal."*

4.5 Perceived Privacy Risks of FR

Privacy is repeatedly brought up as a key concern by our study participants. Around 70% of participants voiced privacy concerns during the study. In this section, we summarize the major themes around perceived privacy risks of facial recognition, in light of concepts from established privacy frameworks (i.e., Solove's "Taxonomy of Privacy" [99] and Westin's states of privacy from *Privacy and Freedom* [109]).

4.5.1 Violation of solitude

The feeling of surveillance prevails A third of our respondents found surveillance through facial recognition to be concerning. Surveillance can exert adverse psychological effects like discomfort and anxiety on subjects. For example, P68 pointed out that *"I had this paranoia that I would be judged based on every action I took at work without the full context."* Similarly, P29 stated that *"always being watched and analyzed which in itself is scary."* Moreover, surveillance is also harmful due to its infringement on people's freedom to act. P89 contextualized this concern — *"There is a feeling of freedom as I enter the library where I participate in a Spanish speaking group on Wednesday morning...in the small classroom where we speak, I would feel rather self-conscious if I were videoed."* This infringement upon freedom can also possibly lead to inhibition and behavior alteration, as P84 noted *"I'd always have to be concerned about how my actions might be perceived on camera,"* and in P20's view, *"I want to know where all of the cameras are, so I can always be aware and I can always be on guard and vigilant. So if*

something happens, I can be ready to defend myself or defend the findings." Surveillance can also have a chilling effect on civil and political engagement. For instance, P117 pointed out that facial recognition *"is used to identify anti-fascists and peaceful protesters"*, and P39 found *"any and all efforts at using such technology against political dissenters"* to be concerning.

Deprived of the right to be let alone Warren and Brandeis first articulated privacy as the "right to be let alone" [108]. Privacy risks also lie in the probing action itself which perturbs this right, making "the person being questioned feel uncomfortable" as noted by Solove [99]. Two-fifths of our participants regarded some deployment scenarios of facial recognition as unwarranted and prying. For instance, P68 manifested their concern, *"It is the idea of somebody being able to surveil my life and know my business...Even though on sight it's something different through a camera, that knowing somebody is interested in the data, and wants it, and is just getting it for free. Something about it really bothers me."* Some participants responded to data collection of facial recognition rather abruptly, *"It's none of anyone's business, as long as I'm obeying the law, where I am and what I'm doing"*(P114). Some participants reported that facial recognition is intrusive into one's life, and they cannot be let alone under the presence of facial recognition. For example, P83 mentioned that they are *"unable to hide from people"*, and P104 noted, *"I feel like I'm being stalked by the man, the powers that be, wealthy corporations."* Others regarded facial recognition as disruptions to their daily activities: P69 mentioned, *"Don't want to be filmed eating,"* and P62 commented on their experience in stores, *"It's like being stared at in the face by someone while I'm just trying to shop."*

4.5.2 Unwanted exposure and violation of anonymity

Not able to stay anonymous 17% of participants stressed the importance of anonymity and scrutinized how facial recognition enabled the identification of normal people in plain view. P63 gave examples of circumstances when people may want to stay anonymous, *"Probably if you go to some kind of clinics, like sexual health clinics, or food pantry."* P12 voiced their concerns about facial recognition used for advertising, *"If it is generating tailored advertising then it implies it is tracking my shopping habits and linking it to my face."* P55 elaborated a situation when he wants to remain anonymous, *"I don't do any sort of very secretive things. The only possible scenarios are if I was trying to...plan a surprise birthday party for my wife, some notification got sent to both of us of where I was, and then she figures that out...There is a mixed scenario of people who are doing slightly illegitimate things but are legal to do, like having affairs with people on their partners, they would definitely not like stuff like that."* Identification, a method to connect people to collected data, is hard to avoid as the deployment of facial recognition technologies becomes widespread.

Unwanted exposure to others This issue involves "exposing to others of certain physical and emotional attributes about a person," which often "creates embarrassment and humiliation" as defined by Solove [99]. 22% of our participants pointed out that it is easy to reveal emotions under contexts of facial recognition involving emotion recognition. For example, P68 described her personal experience, *"I went through a breakup that week. I was really emotional a lot of the time. I do not want my health insurance, my employer, my parents getting updates like "hey, she's trying to get through the pain while she is working today.""* P50 commented on a facial recognition scenario that occurred at the vet they went to, *"People experience deep personal emotions at the vet."* Some respondents were cautious about carrying out private actions. For example, P89 elaborated, *"I might be caught at the gym entering and adjusting a bra strap, etc.,"* and *"doing something like picking your nose, something like that, not doing something against the law, but something you don't want others to see."* Such unwanted exposure in public spaces might not have been feasible without facial recognition technologies.

4.5.3 Non-consensual and insecure disclosure

Secondary use without consent This refers to the privacy issue of data collected for additional purposes without data subjects' knowledge or consent. In the context of facial recognition, this problem is exacerbated because of the lack of ways to properly convey data practices to subjects other than using signs that say "face recognition security cameras in use." Given the sensitive nature of facial recognition data, around a quarter of our participants reported concerns about unauthorized secondary use. Many respondents questioned whether companies would retain data for intended use only, as P12 described, *"As I'm doing this study more, I think it's my trust in their ability to safe keep the data and only for that use. I would doubt their compliance even if I do want them to get the competitive advantage by the use of video surveillance."* P89 also hoped for regulations to prevent secondary use, *"If there were laws in place that they could never ever use it for anything else like they couldn't sell it to marketing companies."* P106 provided a concrete example of secondary use with regards to workplaces using facial recognition to track attendance, *"I think if used to replace a time card is fine, but I could see it being abused by overbearing managers."* A few participants expressed concerns about their data being sold, which can also be regarded as a secondary use.

Fear of data leakage and abuse About one-third of our participants expressed their concerns about their facial recognition data being hacked or abused. Because it is almost im-

practical to relinquish biometric data when compromised, the security of facial recognition data is ever more pressing. Many of the participants reported that they do not trust data collectors' ability to safeguard their data. For example, P122 noted, *"I don't think data security is a strong priority for these companies, and when they do have data leaks, they don't care because it doesn't affect them, and the punishment is not enough to incentivize them to change their practices,"* which parallels the concerns of P54 about identified frivolous activities being leaked, *"Frivolities that end up being insecure, like entertainment or stores."* Also, the fear of insecurity can induce privacy risks by placing people to whom it pertains in a vulnerable state, as corroborated by P122, *"It's very troubling to think of how this info could be used by bad actors."*

4.5.4 Inaccurate dissemination and violation of reserve

Dissemination of inaccurate or misleading information Around one-third of our participants were concerned about the dissemination of inaccurate or misleading information [99]. This issue is also mostly linked to the inaccuracies of facial recognition as presented in Section 4.4.2. Our participants were concerned about being falsely identified or judged out of context. For example, P46 noted, *"Bad luck or timing could lead law enforcement to be suspicious of an innocent citizen."* P11 referred to their experiences when shopping in stores, *"I would really not like supposedly meaningful data to be recorded if I happened to smile remembering something while walking down the condom aisle."* Distortion can be detrimental, as illustrated by P59, *"Reputational damage could occur if someone is falsely accused of a crime."*

Decisional interference Solove defined this as the intrusion on private decisional making, especially by the government [99]. In our study, participants mostly focused on the unwarranted influence on their purchasing autonomy by private companies with the help of facial recognition. This is also discussed in Section 4.3.2. In addition, P89 lamented,*"It's machines taking over and my freedom circumvented."* P122 echoed this thought, *"I do not want to have this information used against me or used to try and subvert my thinking."*

4.6 Proposed Actions and Responses

Our qualitative data also reveals participants' reported desire to take action when encountering facial recognition in their everyday life. They also express a desire for transparency and indicate they would like to be notified about nearby deployments of facial recognition technology. At the same time, their notification preferences vary with some participants expressing concerns about potentially overly disruptive notifications.

4.6.1 Participants want transparency and control over the collection of their data

About 30% of participants expressed strong views about the need for entities collecting sensitive facial recognition data to notify them and to actively obtain consent from them before data collection. For example, P50 commented, *"I think if they are going to record our image, they should have to notify you before they do anything with it like if they are going to use it for a specific purpose, we should be able to know what they are using it for, and we should be able to say "yes, that's fine," or "no, it's not. Delete my stuff from your system.""* While most participants agreed about the need to obtain consent, they did not provide consistent answers with regard to the frequencies of such notifications. Some participants wanted to be notified every time when such data collection is taking place, as illustrated by the quote from P56, *"I think it is important to know when you are in areas where data is being collected, passive consent really disturbs me. I know it happens all the time when I am on my phone or computer, and it is really hard to know what data is being collected, what it is being used for, etc....So, if I have my preference, I would want to know every time someone is engaging in this practice,"* whereas others were wary of repeated reminders and preferred less frequent notices, as P17 elaborated, *"I frequent this establishment pretty often, so a constant reminder would annoy me. It would be nice to be reminded every now and then in case I simply forget."* These results suggest a need for customizable notification functionality where different individuals can select from a number of notification options.

4.6.2 Participants find existing notice mechanisms inadequate

While the majority of participants wanted to be informed about facial recognition in use, our follow-up interviews disclosed the specific ways how some participants found the existing notice mechanisms inadequate. For instance, P68 described how they missed the existing signs in physical spaces that were supposed to notify them about the presence of cameras, *"There will be places where I would want to be notified every time, and then I look over, and see a sign that I have just passed by a dozen times, and realize I am being notified."* When probed about what is a good way to give them notice or obtain their consent, some interviewees reported that no existing mechanisms would achieve the goal, as P53 said, *"I think that [obtaining consent] is hard...It is hard because you cannot pass a form when you walk into a restaurant or a store, it cannot be formal...I guess trying to do it remotely like through the Internet or your phone would be the easiest."* Specifically, P50 expressed their desire to provide consent based on different purposes of facial recognition, *"It would depend on what they were using it for. If it was just like someone committed a crime, and they needed FR for that, then that's fine. Maybe if it's to replace a swipe card or a membership cards, that would*

be okay, but if it's for tracking my purchases, or tracking my attendance, emotions." The information on the purposes for which facial recognition is deployed is not available to data subjects in the majority of current deployments. Also, it is also hard to design notice mechanisms with the desired level of intrusiveness, as P89 elaborated, *"I would not want to think about it at all times, so I want it to be subtle whatever the notification is, but also not so subtle that you don't know that it is happening ever,"* which highlights the problem of privacy as a secondary goal.

4.6.3 Some participants fear being overwhelmed by frequent notifications

While most participants report that they want to be notified, more than half are also weary of too frequent notifications. In particular, some participants realized during the course of the study that the number of notifications they would receive might become a nuisance if they request to be notified each time they get within range of facial recognition technology. For instance, P53 described her thought process, *"When I first started, I was saying once in a while, and then I realized that would be really annoying to get multiple notifications."* About half (55 out of 123) of participants reported that they were unlikely to avoid places that deploy facial recognition technology, even if they indicated being concerned about these deployments, revealing a general sense of resignation. For instance, P11 underscored, *"There is nothing I can do about it, and this is the only accessible grocery around my workplace, so I don't have an alternative."* A similar sense of helplessness and resignation was expressed by P67: *"I give up. Spy on me. What can I do about it? I'm old. I'll be dead soon."*

At the same time, not all participants reported concern. We also observed a small number of participants who did not care about the usage of facial recognition in general, referencing the "nothing to hide" argument. For instance, P55 elaborated, *"I am not likely to be so concerned about it, because I don't do any sort of very secretive things... There are more legitimate reasons why people would want to value their privacy more than I do, but I am not sure how much of the population that would really affect."*

5 Discussion

5.1 Limitations

We would like to first remind the reader that the results presented in this paper focus on a qualitative analysis of data collected as part of our study. A sister publication presented earlier this year provides a quantitative analysis of additional data collected as part of the same experience sampling study [115]. We invite the reader to look at it for additional details about our study protocol and to develop a more comprehensive view of our findings.

We acknowledge that, while ideally, we would have liked to collect data from a representative cross-section of the general public, study participants were recruited from the population of a mid-sized city in the United States (Pittsburgh). Our sample is skewed towards somewhat younger and more educated participants, which might have biased some of our findings. Accordingly, we do not claim that our results are representative of the general population. In addition, our analysis results rely on participants' self-reported qualitative data, which may not necessarily match their actual behaviors.

While describing study scenarios, we strove to maintain a balanced narrative without overly emphasizing benefits or potential risks associated with different deployments. We acknowledge that on occasions, our phrasing might inadvertently have primed participants in one direction or the other.

Finally, our participants generally expressed somewhat negative views of various facial recognition deployment scenarios. This could, in part, be a reflection of the fact that they did not actually experience true interactions with these deployment scenarios and, as a result, may not have had a chance to appreciate what they consider as benefits associated with some of these scenarios (e.g., marketing scenarios).

5.2 Combating Inaccuracy and Bias

While most of participants reported seeing benefits in facial recognition deployments such as security and authentication scenarios, their reported attitude towards many other scenarios was generally more negative. Part of their willingness to embrace the technology was dampened by concerns over accuracy and bias of facial recognition systems, echoing concerns voiced by marginalized interviewees in a prior study [47]. Our data suggest that these concerns extend to the more general population. Recent reports of people wrongly arrested due to faulty facial recognition algorithms likely contributed to reservations captured in our study [52] and also illustrate the severe consequences that deployment of this technology can have if deployed and relied upon without adequate safeguards. Minimally, technology should be evaluated for potential biases and minimal levels of accuracy, especially when deployed in support of particularly sensitive activities such as law enforcement. Their performance and limitations should be clearly communicated and taken into account. And decisions based on these algorithms should be meticulously cross-checked and manually vetted if we are to avoid more of these nightmarish scenarios.

5.3 Contextualizing Perceived Privacy Risks

Our analysis organized perceived privacy risks associated with facial recognition deployments around key dimensions identified in well-established privacy frameworks [99, 108, 109].We were able to elicit more nuanced and contextualized privacy concerns than prior work [21, 97, 100] as shown in

Section 4.5. While legal arguments support people's reasonable expectations of privacy in public places [53], our study provides strong evidence that these expectations are real and widespread and that some facial recognition deployment scenarios are perceived as overstepping the boundaries of personal solitude, making people feel deprived of "the(ir) right to be let alone" [108]. These concerns are further exacerbated by the sensitive nature of biometric data, the information that can be inferred from facial recognition data (e.g., location, activity, and mood), as well as risks of secondary use of this data and its security. These findings underscore the need for more transparency in notifying people about not just the deployment of facial recognition technology but also sufficient details for individuals to gauge their perceived privacy risks.

5.4 Designing Effective Notice and Choice

Our study confirms that privacy concerns are a major obstacle to acceptance of a variety of facial recognition scenarios [21, 22, 83], although these deployments are becoming increasingly widespread. Responses from our participants indicate a strong desire to be notified about different deployment scenarios and to have some control over the collection and analysis of their data. Current deployments generally fall short when it comes to effectively notifying people about the presence of facial recognition technologies, including details about the type of analysis they rely on and how results are being used and possibly shared. Also, current deployments generally fail to provide people with opt-in or opt-out choices.

How to effectively notify people and offer them adequate controls is not trivial. Entities deploying facial recognition should inform data subjects in a clear and noticeable manner. Today's "this area under camera surveillance" signs do not provide them with enough information, such as type of analysis, the purpose for collection and analysis, sharing, etc. Privacy controls (e.g., opt-in and opt-out choices) should obviously include mechanisms to authenticate data subjects (to make sure they are whom they claim to be when they request to opt in or out of some practices), giving rise to privacy issues. With the possible exception of security-related deployments, which many view as generally beneficial, people should be offered some control over the collection and use of their footage — preferably in the form of opt-ins.

One solution involves requiring people to opt in by providing training data about their face [27, 28]. In this system, a privacy-aware infrastructure is used to notify people about the presence of nearby facial recognition deployments, including who has deployed the technology, what analysis is performed, and for how long the footage is retained. Users who do not opt in for facial recognition by default have their face (or possibly their entire body) obfuscated in real-time in the captured footage. Notifications about nearby facial recognition deployment are provided via a "Privacy Assistant" mobile app that users install on their smartphones. This infrastructure has been deployed to support notice and choice for a variety of Internet of Things data collection processes — not just facial recognition [27, 88].

Our data highlight individuals' diverse notification preferences, with some preferring to be systematically notified about FR deployments, while others only would prefer just occasional notices and reminders. The Internet of Things Privacy Infrastructure introduced by Das et al. offers users of its "Privacy Assistant" mobile app different settings they can configure to specify the types of data collection processes they want to be notified about as well as the frequency of these notifications (e.g., "only the first time," "every time," or "never"). These settings are consistent with results discussed in Section 4.6, which indicate that different participants have different notification preferences and that these preferences can also evolve. Further research is needed to determine what personalized settings are likely to work best and how to alleviate the user burden that might be entailed by opt-in or opt-out settings associated with a potentially large number of facial recognition deployments.

Finally, our study indicates that participants fear losing their autonomy when commercial entities can assemble and leverage near real-time facial recognition data, including their emotions, to tailor advertisements presented to them. Our participants also expressed reservations about the power this technology can bestow on already powerful entities such as their employers or law enforcement authorities. These results further emphasize the need for more effective notice and choice mechanisms if people are to become less fearful about the deployment of facial recognition.

6 Conclusion

Deployment of facial recognition technologies is already widespread and continuing to grow. While many people are familiar with typical video surveillance scenarios, most have little or no awareness of the increasingly diverse set of scenarios where this technology is being deployed. We analyzed data from a 10-day in-situ study where we collected information about people's awareness and perceptions of a variety of facial recognition deployments they could realistically encounter as part of their everyday activities. Our data show that people's privacy concerns are complex and depend on different attributes characterizing these deployment scenarios. Our analysis reveals serious concerns about the privacy impact of these technologies, including the lack of mechanisms to effectively notify people and give them some control over the collection, analysis, and use of their footage. Our data also suggest that people's views about facial recognition technologies have been impacted by recent reports about the inconsistent accuracy and bias found in deployed systems. The qualitative analysis presented in this paper complements a quantitative analysis of data collected as part of the same study presented in a recent sister publication [115].

Acknowledgments

We thank Dr. Xu Wang and Zheng Yao for their input on this paper. We also thank Dr. Lujo Bauer, Dr. Lorrie Cranor, and Dr. Anupam Das for their feedback on the study. This research has been supported in part by DARPA and AFRL under agreement number FA8750-15-2-0277 and in part by NSF under grants from the National Science Foundation Secure and Trustworthy Computing program (CNS-15-13957, CNS-1801316). The US Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright notice thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies or endorsements, either expressed or implied of DARPA, AFRL, NSF or the US Government.

References

[1] Augmented mental health: Revolutionary mental health care using emotion recognition. https://www.augmentedmentalhealth.com/blog/augmented-mental-health-revolutionary-mental-health-care-using-emotion-recognition, 2018.

[2] Chinese man caught by facial recognition at pop concert. https://www.bbc.com/news/world-asia-china-43751276, 2018.

[3] Facial recognition: School ID checks lead to GDPR fine. https://www.bbc.com/news/technology-49489154, 2019.

[4] Facial recognition technology: Ensuring transparency in government use. https://www.nist.gov/speech-testimony/facial-recognition-technology-ensuring-transparency-government-use, 2019.

[5] ACM U.S. Technology Policy Committee. Statement on principles and prerequisites for the development, evaluation and use of unbiased facial recognition technologies. https://www.acm.org/binaries/content/assets/public-policy/ustpc-facial-recognition-tech-statement.pdf, 2020.

[6] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:2037–2041, 2006.

[7] Marshall Allen. Health insurers are vacuuming up details about you — and it could raise your rates. https://www.propublica.org/article/health-insurers-are-vacuuming-up-details-about-you-and-it-could-raise-your-rates, 2018.

[8] Nazanin Andalibi and Justin Buss. The human in emotion recognition on social media: Attitudes, outcomes, risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–16, New York, NY, USA, 2020. ACM.

[9] Association Press. Tesco's plan to tailor adverts via facial recognition stokes privacy fears. https://www.theguardian.com/business/2013/nov/03/privacy-tesco-scan-customers-faces, 2013.

[10] Rachel Bachman. Your gym's tech wants to know you better. https://www.wsj.com/articles/your-gyms-tech-wants-to-know-you-better-1497281915, 2017.

[11] Sarah Pulliam Bailey. Skipping church? Facial recognition software could be tracking you. http://www.washingtonpost.com/news/acts-of-faith/wp/2015/07/24/skipping-church-facial-recognition-software-could-be-tracking-you/, 2015.

[12] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M Martinez, and Seth D Pollak. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. *Psychol Sci Public Interest.*, 20(3):165–166, 2019.

[13] Daniel J Beal. ESM 2.0: State of the art and future potential of experience sampling methods in organizational research. *Annual review of organizational psychology and organizational behavior*, 2(1):383–407, 2015.

[14] Peter N. Belhumeur, João P. Hespanha, and David J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:711–720, 1997.

[15] Bloomberg News. Mannequins collect data on shoppers via facial-recognition software. https://www.washingtonpost.com/business/economy/mannequins-collect-data-on-shoppers-via-facial-recognition-software/2012/11/22/0751b992-3425-11e2-9cfa-e41bac906cc9_story.html, 2012.

[16] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[17] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

[18] David Burrows. Facial expressions show Mars the adverts that will drive sales. https://www.foodnavigator.com/Article/2017/03/23/Facial-expressions-show-Mars-the-adverts-that-will-drive-sales, 2017.

[19] Ramon Caceres and Adrian Friday. Ubicomp systems at 20: Progress, opportunities, and challenges. *IEEE Pervasive Computing*, 11(1):14–21, 2011.

[20] Laura L. Carstensen et al. Emotional experience improves with age: Evidence based on over 10 years of experience sampling. *Psychology and Aging*, 26(1):21, 2011.

[21] Daniel Castro and McLaughlin Michael. Survey: Few Americans want government to limit use of facial recognition technology, particularly for public safety or airport screening. https://datainnovation.org/2019/01/survey-few-americans-want-government-to-limit-use-of-facial-recognition-technology-particularly-for-public-safety-or-airport-screening/, 2019.

[22] Richard Chow. The last mile for IoT privacy. *IEEE Security & Privacy*, 15(6):73–76, 2017.

[23] Ben Conarck. Florida court: Prosecutors had no obligation to turn over facial recognition evidence. https://www.jacksonville.com/news/20190123/florida-court-prosecutors-had-no-obligation-to-turn-over-facial-recognition-evidence, 2019.

[24] Kate Conger, Richard Fausset, and Serge F. Kovaleski. San Francisco bans facial recognition technology. https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html, 2019.

[25] Elly Cosgrove. One billion surveillance cameras will be watching around the world in 2021, a new study says. https://www.cnbc.com/2019/12/06/one-billion-surveillance-cameras-will-be-watching-globally-in-2021.html, 2019.

[26] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-FCN: Object detection via region-based fully convolutional networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 379–387, Red Hook, NY, USA, 2016. Curran Associates Inc.

[27] Anupam Das, Martin Degeling, Daniel Smullen, and Norman Sadeh. Personalized privacy assistants for the Internet of Things: Providing users with notice and choice. *IEEE Pervasive Computing*, 17(3):35–46, 2018.

[28] Anupam Das et al. Assisting users in a world full of cameras: A privacy-aware infrastructure for computer vision applications. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1387–1396. IEEE, 2017.

[29] Bobby J Davidson. How your business can benefit from facial recognition technology. https://percentotech.com/how-your-business-can-benefit-from-facial-recognition-technology/, 2019.

[30] Dean DeChiaro. New York City eyes regulation of facial recognition technology. https://www.rollcall.com/news/congress/new-york-city-eyes-regulation-of-facial-recognition-technology, 2019.

[31] Benchaa Djellali, Kheira Belarbi, Abdallah Chouarfia, and Pascal Lorenz. User authentication scheme preserving anonymity for ubiquitous devices. *Security and Communication Networks*, 8(17):3131–3141, 2015.

[32] Yitao Duan and John Canny. Protecting user data in ubiquitous computing: Towards trustworthy environments. In *International Workshop on Privacy Enhancing Technologies*, pages 167–185. Springer, 2004.

[33] Melanie Ehrenkranz. Burger joint teams up with surveillance giant to scan your face for loyalty points. https://gizmodo.com/burger-joint-teams-up-with-surveillance-giant-to-scan-y-1821498988, 2017.

[34] Zekeriya Erkin et al. Privacy-preserving face recognition. In Ian Goldberg and Mikhail J. Atallah, editors, *Privacy Enhancing Technologies*, pages 235–253, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[35] Darrell Etherington. Baidu and KFC's new smart restaurant suggests what to order based on your face. https://techcrunch.com/2016/12/23/baidu-and-kfcs-new-smart-restaurant-suggests-what-to-order-based-on-your-face/, 2016.

[36] Ingrid Fadelli. Analyzing spoken language and 3-D facial expressions to measure depression severity. https://techxplore.com/news/2018-11-spoken-language-d-facial-depression.html, 2019.

[37] Caitlin Fairchild. Hertz is now using facial recognition to check out cars. https://www.nextgov.com/emerging-tech/2018/12/hertz-now-using-facial-recognition-check-out-cars/153479/, 2018.

[38] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[39] Denzil Ferreira, Jorge Goncalves, Vassilis Kostakos, Louise Barkhuus, and Anind K. Dey. Contextual experience sampling of mobile application micro-usage. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services (MobileHCI '14)*, pages 91–100, 2014.

[40] Chris Frey. Revealed: how facial recognition has invaded shops—and your privacy. https://www.theguardian.com/cities/2016/mar/03/revealed-facial-recognition-software-infiltrating-cities-saks-toronto, 2016.

[41] Sarah Fister Gale. Employers turn to biometric technology to track attendance. https://www.workforce.com/news/employers-turn-to-biometric-technology-to-track-attendance, 2013.

[42] Shirin Ghaffary and Rani Molla. Here's where the us government is using facial recognition technology to surveil Americans. https://www.vox.com/recode/2019/7/18/20698307/facial-recognition-technology-us-government-fight-for-the-future, 2019.

[43] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Ongoing face recognition vendor test (FRVT) part 2: Identification. https://nvlpubs.nist.gov/nistpubs/ir/2018/NIST.IR.8238.pdf, 2018.

[44] Patrick Grother, Mei Ngan, and Kayee Hanaoka. Ongoing face recognition vendor test (FRVT) part 1: Verification. https://www.nist.gov/system/files/documents/2019/11/20/frvt_report_2019_11_19_0.pdf, 2019.

[45] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. MS-Celeb-1M: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016.

[46] Yaron Gurovich et al. Identifying facial phenotypes of genetic disorders using deep learning. *Nature Medicine*, 25(1):60–64, 2019.

[47] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. *Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems*, page 1–13. ACM, New York, NY, USA, 2018.

[48] Drew Harwell. Amazon extends ban on police use of its facial recognition technology indefinitely. https://www.washingtonpost.com/technology/2021/05/18/amazon-facial-recognition-ban/, 2021.

[49] Drew Harwell. Senators seek limits on some facial-recognition use by police, energizing surveillance technology debate. https://www.washingtonpost.com/technology/2021/04/21/data-surveillance-bill/, 2021.

[50] Xiaofei He, Shuicheng Yan, Yuxiao Hu, Partha Niyogi, and Hong-Jiang Zhang. Face recognition using Laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:328–340, 2005.

[51] Joel M. Hektner, Jennifer A. Schmidt, and Mihaly Csikszentmihalyi. *Experience sampling method: Measuring the quality of everyday life*. Sage, 2007.

[52] Kashmir Hill. Wrongfully accused by an algorithm. https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html, 2020.

[53] Mariko Hirose. Privacy in public spaces: The reasonable expectation of privacy against the dragnet use of facial recognition technology. *Connecticut Law Review*, (377), 2017.

[54] Wilhelm Hofmann, Roy F. Baumeister, Georg Förster, and Kathleen D. Vohs. Everyday temptations: An experience sampling study of desire, conflict, and self-control. *Journal of Personality and Social Psychology*, 102(6):1318, 2012.

[55] Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. 2008.

[56] Isabelle Hupont and Carles Fernández. DemogPairs: Quantifying the impact of demographic imbalance in deep face recognition. *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–7, 2019.

[57] Timothy Johnson. Shoplifters meet their match as retailers deploy facial recognition cameras. https://www.mcclatchydc.com/news/nation-world/national/article211455924.html, 2018.

[58] Ira Kemelmacher-Shlizerman, Steven M. Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4873–4882, 2016.

[59] Os Keyes. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), 2018.

[60] Mehreen Khan. EU plans sweeping regulation of facial recognition. https://www.ft.com/content/90ce2dce-c413-11e9-a8e9-296ca66511c9, 2019.

[61] Ingrid Kramer et al. A therapeutic application of the experience sampling method in the treatment of depression: a randomized controlled trial. *World Psychiatry*, 13(1):68–77, 2014.

[62] Sarah Krouse. The new ways your boss is spying on you. https://www.wsj.com/articles/the-new-ways-your-boss-is-spying-on-you-11563528604, 2019.

[63] Stephen Lepitak. Disney's Dumbo and Accenture Interactive collaborate for the movie poster of the future. https://www.thedrum.com/news/2019/03/10/disneys-dumbo-and-accenture-interactive-collaborate-the-movie-poster-the-future, 2019.

[64] Gil Levi and Tal Hassner. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, page 503–510, New York, NY, USA, 2015. ACM.

[65] David Levine. What high-tech tools are available to fight depression? https://health.usnews.com/health-care/patient-advice/articles/2017-10-06/what-high-tech-tools-are-available-to-fight-depression, 2017.

[66] David Levine. What your face may tell lenders about whether you're creditworthy. https://www.wsj.com/articles/what-your-face-may-tell-lenders-about-whether-youre-creditworthy-11560218700, 2019.

[67] Chengjun Liu and Harry Wechsler. Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image Processing*, 11 4:467–76, 2002.

[68] Brain Logan. Pay-per-laugh: the comedy club that charges punters having fun. https://www.theguardian.com/stage/2014/oct/14/standup-comedy-pay-per-laugh-charge-barcelona, 2014.

[69] Martin Magdin, L'ubomír Benko, and Štefan Koprda. A case study of facial emotion classification using Affdex. *Sensors (Basel)*, 19(9):2140, 2019.

[70] Tobias Matzner. Why privacy is not enough privacy in the context of "ubiquitous computing" and "big data". *Journal of Information, Communication and Ethics in Society*, 12(2):93–106, 2014.

[71] Darren Murph. SceneTap app analyzes pubs and clubs in real-time, probably won't score you a Jersey Shore cameo. https://www.engadget.com/2011/06/12/scenetap-app-analyzes-pubs-and-clubs-in-real-time-probably-won/, 2011.

[72] Sharon Nakar and Dov Greenbaum. Now you see me. Now you still do: Facial recognition technology and the growing lack of privacy. *Boston University Journal of Science & Technology Law*, (23):88–122, 2017.

[73] NEC Corporation. New biometric identification tools used in theme parks. https://www.nec.com/en/global/about/mitatv/03/3.html, 2002.

[74] Alfred Ng. With facial recognition, shoplifting may get you banned in places you've never been. https://www.cnet.com/news/with-facial-recognition-shoplifting-may-get-you-banned-in-places-youve-never-been/, 2019.

[75] Helen Nissenbaum. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, 2009.

[76] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015.

[77] PCMag Stuff. NEC unveils facial-recognition system to identify shoppers. https://www.pcmag.com/archive/nec-unveils-facial-recognition-system-to-identify-shoppers-305015, 2012.

[78] Charith Perera, Rajiv Ranjan, Lizhe Wang, Samee U Khan, and Albert Y. Zomaya. Big data privacy in the Internet of Things era. *IT Professional*, 17(3):32–39, 2015.

[79] Salil Prabhakar, Sharath Pankanti, and Anil K. Jain. Biometric recognition: Security and privacy concerns. *IEEE Security & Privacy*, 1(2):33–42, 2003.

[80] Emilee Rader. Most Americans don't realize what companies can predict from their data. https://bigthink.com/technology-innovation/most-americans-dont-realize-what-companies-can-predict-from-their-data-2629911919, 2019.

[81] Inioluwa D. Raji et al. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, pages 145–151, New York, NY, USA, 2020. ACM.

[82] Edith Ramirez, Julie Brill, Maureen K. Ohlhausen, Joshua D. Wright, and Terrell McSweeny. Data brokers: A call for transparency and accountability. Technical report, Federal Trade Commission, 2014.

[83] Luis Felipe M. Ramos. Evaluating privacy during the covid-19 public health emergency: The case of facial recognition technologies. In *Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance*, ICEGOV 2020, page 176–179, New York, NY, USA, 2020. ACM.

[84] Robert W. Reeder et al. An experience sampling study of user reactions to browser warnings in the field. In

Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pages 1–13, 2018.

[85] Timothy Revell. Computer vision algorithms pick out petty crime in CCTV footage. https://www.newscientist.com/article/2116970-computer-vision-algorithms-pick-out-petty-crime-in-cctv-footage/, 2017.

[86] David Rosen. Disney is spying on you! https://www.salon.com/test/2013/01/17/disney_is_spying_on_you/, 2013.

[87] Andrew Ryan et al. Automated facial expression recognition system. In *43rd Annual 2009 International Carnahan Conference on Security Technology*, pages 172–177, 2009.

[88] Norman Sadeh. Design of a privacy infrastructure for the internet of things. In *2020 USENIX Conference on Privacy Engineering Practice and Respect (PEPR 20)*. USENIX Association, 2020.

[89] T. Scott Saponas, Jonathan Lester, Carl Hartung, Sameer Agarwal, and Tadayoshi Kohno. Devices that tell on you: Privacy trends in consumer ubiquitous computing. In *16th USENIX Security Symposium (USENIX Security '07)*, pages 55–70, 2007.

[90] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[91] E. J. Schultz. Facial-recognition lets marketers gauge consumers' real responses to ads. https://adage.com/article/digital/facial-recognition-lets-marketers-gauge-real-responses/298635, 2015.

[92] Christie N. Scollon, Chu Kim-Prieto, and Ed Diener. Experience sampling: Promises and pitfalls, strengths and weaknesses. *Journal of Happiness Studies*, 4(1):5–34, 2003.

[93] Ignacio Serna et al. SensitiveLoss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. *arXiv*, abs/2004.11246, 2020.

[94] Shawn Shan et al. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX Security Symposium (USENIX Security '20)*, pages 1589–1604, 2020.

[95] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS '16)*, pages 1528–1540, 2016.

[96] Ed Silverstein. New Konami casino facial recognition technology could rival reward cards. https://www.casino.org/news/new-konami-casino-facial-recognition-technology-could-rival-reward-cards/, 2019.

[97] Arron Smith. More than half of U.S. adults trust law enforcement to use facial recognition responsibly. Technical report, Pew Research Center, 2019.

[98] Benjamin Snyder. This beer ad only works when women pass by. https://fortune.com/2015/05/21/astra-beer-ad/, 2015.

[99] Daniel J. Solove. A taxonomy of privacy. *University of Pennsylvania Law Review*, 154(3):477–564, 2006.

[100] Luke Stark, Amanda Stanhaus, and Denise L. Anthony. "I don't want someone to watch me while I'm working": Gendered views of facial recognition technology in workplace surveillance. *Journal of the Association for Information Science and Technology*, 71(9):1074–1088, 2020.

[101] Léa Steinacker, Miriam Meckel, Genia Kostka, and Damian Borth. Facial recognition: A cross-national survey on public acceptance, privacy, and discrimination. In *International Conference on Machine Learning - Law and ML Workshop*, 2020.

[102] Steve Stemler. An overview of content analysis. *Practical assessment, research, and evaluation*, 7(1):17, 2000.

[103] Francesca Street. How facial recognition is taking over airports. https://www.cnn.com/travel/article/airports-facial-recognition/index.html, 2019.

[104] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014.

[105] U.S. Government Accountability Office. Facial recognition technology: Commercial uses, privacy issues, and applicable federal law. https://www.gao.gov/products/GAO-15-621, 2015.

[106] Niels Van Berkel, Denzil Ferreira, and Vassilis Kostakos. The experience sampling method on mobile devices. *ACM Computing Surveys*, 50(6):1–40, 2017.

[107] Simone J. W. Verhagen, Laila Hasmi, Marjan Drukker, Jim van Os, and Philippe A. E. G. Delespaul. Use of the experience sampling method in the context of clinical trials. *Evidence-based Mental Health*, 19(3):86–89, 2016.

[108] Samuel D. Warren and Louis D. Brandeis. The right to privacy. *Harvard Law Review*, 4(5):193–220, 1890.

[109] Alan F. Westin. Privacy and freedom. *Washington & Lee Law Review*, 25:166, 1968.

[110] Jason Whitely. How facial recognition technology is being used, from police to a soccer museum. https://www.wfaa.com/article/features/originals/how-facial-recognition-technology-is-being-used-from-police-to-a-soccer-museum/287-618278039, 2018.

[111] Niels Wouters et al. Biometric mirror: Exploring ethical opinions towards facial analysis and automated decision-making. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, DIS '19, page 447–461, New York, NY, USA, 2019. ACM.

[112] Elisa Wright. The future of facial recognition is not fully known: Developing privacy and security regulatory mechanisms for facial recognition in the retail sector. *Fordham Intellectual Property, Media & Entertainment Law Journal*, 29(2):611–685, 2019.

[113] John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:210–227, 2009.

[114] Huijuan Xu, Abir Das, and Kate Saenko. R-C3D: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[115] Shikun Zhang et al. "Did you know this camera tracks your mood?": Understanding privacy expectations and preferences in the age of video analytics. *Proc. Priv. Enhancing Technol.*, 2021(2):282–304, 2021.

[116] Shikun Aerin Zhang et al. Understanding people's privacy attitudes towards video analytics technologies. Technical Report CMU-ISR-20-114, Carnegie Mellon University, School of Computer Science, 2020.

7 Appendix

7.1 Evening Review

[Show a map, timestamp and scenario for each notification]

- We asked: How surprised would you be about [PLACE] engaging in this data practice? At the time, you indicated that you would find this ___. Why?
- We asked: How comfortable would you feel about [PLACE] engaging in this data practice? At the time, you indicated that you would find this ___. Why?
- We asked: How would you want to be notified as you enter [PLACE]? At the time, you indicated that you _____. Why?
- If you had the choice, would you allow or deny this data practice?
- Based on the data practice description above, do you believe the footage in which you appear could be made available to third parties for analysis with facial recognition?
- Please indicate how much you agree or disagree with each of the following statements.
  - I feel that I benefit from this data practice
  - I feel that [PLACE] benefits from this data practice
  - I feel that the data practice enhances public safety
- How would you feel about the raw footage being shared with the following entities?

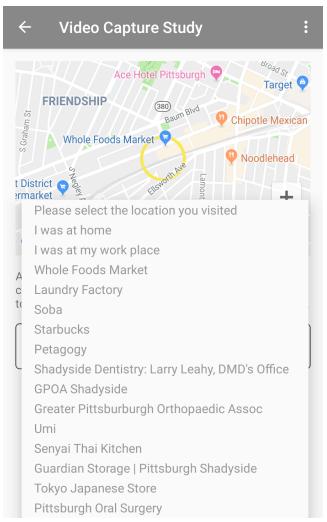
7.2 Post Study Survey

- What is the first thing that comes to your mind when you think about facial recognition technology?
- In what context(s) do you find the use of facial recognition technology to be particularly beneficial? (Enter up to 5 types of contexts)
- In what context(s) do you find the use of facial recognition technology to be particularly concerning? (Enter up to 5 types of contexts)
- Do you feel that you have a general understanding of where this type of technology is likely to be used and why?
- Please rate your comfort level when visiting stores and other locations that use facial recognition technology.
- How likely would you be to intentionally avoid stores that use facial recognition technology?
- Has your level of concern about facial recognition technology changed over the course of the study?
- 10 IUIPC Questions
- Show scenarios: Petty Crime/Sentiment Ads(IDed)/Health Predictions
  - Within what timeframe, do you believe this data practice will be commonplace?
  - Would you like to be notified about this data practice?
  - What sensitive information do you think could be inferred from this data collection practice?
  - How concerned would you be about this sensitive information being inferred? Why?
  - How likely would you be to avoid visiting those places following the introduction of this data practice?
  - What do you think is a reasonable timeframe for those places to retain the footage they capture of you?
  - In what manner would you like to receive notification about those places' use of this data practice?

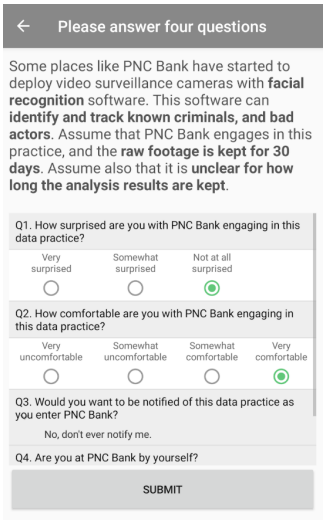
7.3 Interview Scripts

- Introduction: Thank you for agreeing to this interview. My name is ___. I will be audio-recording our session. How are you doing today? Just to fresh your memory. You started the study around [DATE], and finished the study around [DATE]. For this interview, we will be asking you some additional questions and clarifications about your experience during this study.
- Where did you find about our study?
- When did you download the app?
- Did you find participating in this study to be demanding?
- On average, how much time would you say you spent answering our questions each day?
- Were there days when you didn't receive any prompts?
- On the whole, do you feel that we covered most of the interesting places you went to during the course of the study, or would you say we missed some interesting places? If so, which interesting places did we miss? Would you expect cameras to be present at these places and what do you think these cameras could be doing?
- While going through the evening reviews, did you ever feel that you wished you could modify some of the answers you provided during the day? If so, can you specifically remember some of the scenarios and in which way you would have liked to modify your answers (e.g., less surprised or more surprised, less comfortable or more comfortable?)
- [CHECK DATA] For scenarios where we only collected your answers in the evening, because you didn't have time to answer them when the scenario occurred. Do you believe that you might have given different answers if you had responded at the time we first prompted you? If so, how different would your answer have been and why?
- [SHOW INSTANCES] Do you remember when you did not answer those scenarios on site / in-situ, why you could not answer them, and what you were you doing at the time?
- If you remember, each scenario came with two questions designed to check whether you had carefully read the description of the scenario. Do you remember those?
- Did you find that answering these questions could easily be defeated, or did you actually have to carefully read the scenarios to answer the questions? Feel free to tell us that the questions were easy to guess without reading the scenarios. We are trying to understand to what extent these questions help, or to what extent they are just not terribly useful.
- How often did you think that the scenarios we described matched actual video collection practices at the places you were visiting?
- Did you actually look for cameras, or start paying more attention to cameras?
- Have you discussed the study or scenarios with others?
- On the whole, do you feel that you have grown more concerned or less concerned about the types of video analytics scenarios used in our study? Or would you have you remain equally concerned or unconcerned?
- If you remembered, there are a lot of scenarios you encountered as part of this study, were there scenarios that you found particularly surprising? Were there some scenarios that you found particular concerning? Or would you say that all these scenarios are to be expected and do not feel particularly concerning?
- Do you feel that, if you were to retake the study and be presented with the same scenarios, most of your answers would be the same? If some of your answers are likely to be different, could you identify some of the scenarios for which you would likely have different answers?
- Questions/Clarifications related to the interviewee's post-study and evening answers (different case by case)

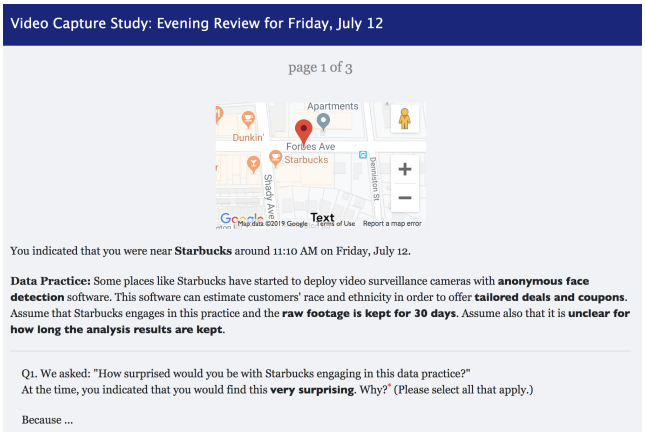
7.4 Screenshots



(a) Users clarify their location



(b) In-situ scenarios



(c) Partial screenshot of the evening survey associated with a given scenario encountered earlier during the day

Figure 1: Screenshots of study instruments

7.5 Scenario Texts

Purpose	Scenario Text
Generic Surveillance	Some places like %s have started to deploy video surveillance cameras to deter crime.
Petty Crime	Some places like %s have started to deploy video surveillance cameras to deter crime. These cameras are equipped with software that can automatically detect and record petty crime (e.g. pickpocketing, car break-ins, breaking store windows).
Known Criminal Detection	Some places like %s have started to deploy video surveillance cameras with facial recognition software. This software can identify and track known shoplifters, criminals, and bad actors.
Count people	Some places like %s have started to deploy video surveillance cameras with anonymous face detection software. This software can estimate the number of customers in the facility in order to optimize operation, such as personnel allocation.
Jump Line	Some places like %s have started to deploy video surveillance cameras with facial recognition software. This software can identify patrons in line and push individualized offers to skip the wait-line for a fee.
Targeted Ads(Anon)	Some places like %s have started to deploy video surveillance cameras with anonymous face detection software. This software can estimate customers' race and ethnicity in order to offer tailored deals and coupons.
Targeted Ads(IDed)	Some places like %s have started to deploy video surveillance cameras with facial recognition software. This software can match detected faces against individual customer profiles in order to offer tailored deals and coupons.
Sentiment Ads(Anon)	Some places like %s have started to deploy video surveillance cameras with anonymous face detection and emotion analysis software. This software can estimate customers' age, gender and ethnicity, and analyze their reactions to items displayed. This software is used to generate tailored deals and coupons for different demographic groups.
Sentiment Ads(IDed)	Some places like %s have started to deploy video surveillance cameras with facial recognition and emotion analysis software. This software can recognize people, and analyze their reactions to items displayed. Then the software matches detected faces against individual customer profiles to send tailored deals and coupons to their phones.
Rate Service	Some places like %s have started to deploy video surveillance cameras with anonymous emotion analysis software. This software can gauge customer satisfaction with the service provided by its employees. They can use the results for employee evaluation and training purposes.
Rate Engagement	Some places like %s have started to deploy video surveillance cameras with facial recognition and emotion analysis software. This software can identify each patron, and measure their engagement at the facility.
Face as ID	Some places have started to deploy video surveillance cameras with facial recognition software. This software can identify faces to replace ID cards.
Track Attendance	Some companies have started to deploy video surveillance cameras with facial recognition software. This software can track the work time attendance of its employees.
Word Productivity	Some companies have started to deploy video surveillance cameras with emotion analysis and facial recognition software. This software can detect the mood of its employees, and predict their productivity. This software can record your presence and who you hang out with.
Health Predictions	Some eatery chains like %s have started to deploy video surveillance cameras with emotion analysis and facial recognition software. This software can detect your mood, and record data about your orders. This information can be shared with health insurance providers. The health insurance providers could use such data to estimate your likelihood of developing depression, diabetes, and obesity, which in turn can impact your health insurance premium.
Medical Predictions	Some medical facilities have started to deploy video surveillance cameras with emotion analysis and facial recognition software. This software can automatically detect some physical and mental health problems. This information can be shared with health insurance providers, which could impact your health insurance premium.

Table 4: Scenarios text shown to participants

"I'm Literally Just Hoping This Will Work:" Obstacles Blocking the Online Security and Privacy of Users with Visual Disabilities

Daniela Napoli, Khadija Baig, Sana Maqsood, Sonia Chiasson
Carleton University
{daniela.napoli, khadija.baig, sana.maqsood}@carleton.ca, chiasson@scs.carleton.ca

Abstract

To successfully manage security and privacy threats, users must be able to perceive the relevant information. However, a number of accessibility obstacles impede the access of such information for users with visual disabilities, and could mislead them into incorrectly assessing their security and privacy. We explore how these users protect their online security and privacy. We observed their behaviours when navigating Gmail, Amazon, and a phishing site imitating CNIB, a well-known organization for our participants. We further investigate their real world concerns through semi-structured interviews. Our analysis uncovered severe usability issues which led users to engage in risky behaviours or to compromise between accessibility or security. Our work confirms the findings from related literature and provides novel insights, such as how software for security (e.g., antivirus) and accessibility (e.g., JAWS) can hinder users' abilities to identify risks. We organize our main findings around four states of security and privacy experienced by users while completing sensitive tasks, and provide design recommendations for communicating security and privacy information to users with visual disabilities.

1 Introduction

More than 2 billion people worldwide live with some form of visual disability [36]. In this paper, we work with individuals with limited visual function such as those who are blind, have low vision, or have other visual disabilities. These individuals' visual capabilities are not situational nor can be changed with corrective lenses. Accessible technologies allow people with visual disabilities to autonomously achieve tasks, which improves their overall quality of life [15].

In practice, they often encounter usability issues, even when services meet common accessibility guidelines [6, 37, 44]. Examples include: confusing or misleading feedback, insufficient information, and compatibility issues between operating systems and assistive software [9, 29, 47]. As such, accessibility and usability are interdependent, emphasizing that approaching web accessibility in isolation is ineffective.

Practical guidelines and frameworks for user-centered security have been proposed [17, 18, 27, 35, 50], but most of the proposed solutions for managing web-based threats are visual which makes them inaccessible to users with visual disabilities, thereby compromising their security and privacy.

Prior work on the security and privacy concerns of users with visual disabilities [3, 24, 25, 48], highlights the unique challenges of designing accessible security and privacy systems. Specifically, that it requires designers to carefully consider and implement both accessibility and usable security design guidelines. Our work adds to this growing body of literature by focusing on users with visual disabilities' security and privacy experiences while web browsing in situations where they are working with potentially sensitive information.

To this end, we conducted a task-based user study and semi-structured interviews with 14 users to identify their security and privacy concerns while web browsing and the effectiveness of their protection strategies. Our work was guided by the following research questions:

RQ1: What types of online security/privacy concerns and barriers exist for those with visual disabilities when visiting websites?

RQ2: Are web security cues accessible and can they be easily interpreted?

RQ3: How do users with visual disabilities perceive and manage web-based risks and threats?

Based on both the study tasks and interviews about real-life practices, we found that users with visual disabilities experienced a number of severe security and privacy-related issues

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

including: inaccessible antivirus software, misleading screen reader outputs, and ill-fitting security advice. These issues can lead to increased security and privacy risks for users. For example, none of our participants were able to correctly identify our phishing website as potentially malicious.

From our findings, we identify four security and privacy awareness "states" which consider accessibility challenges and their influence on security and privacy strategies over time. We propose design recommendations which better suit the capabilities of people with visual disabilities.

2 Background

Many users with visual disabilities routinely complete transactions (e.g., banking and shopping) online, but face severe accessibility issues and have privacy or security concerns [25, 42]. Their major concerns include viruses, encountering CAPTCHAs, spam emails, unauthorized access to search history, and location-based data tracking. Some of these concerns could be addressed through specialized security software.

However, existing security software is often inaccessible and incompatible with screen reader keyboard short-cuts [42]. Similarly, Dosono et al. [19] observed 12 users with visual disabilities use email, banking, and eCommerce websites via screen readers. Poorly labelled login elements confused users with visual disabilities. Additionally, other accessibility issues with audible password masking, insufficient error messages, and password recovery methods negatively impacted users' control of their accounts containing sensitive information.

Ahmed et al. [3] interviewed 14 users with visual disabilities and found that privacy issues forced them to rely on inconvenient workarounds like disabling the screen (even if they require visual cues), wearing headphones (minimising their awareness of physical surroundings), and relying on sighted assistants to complete transactions on their behalf. Hayes et al. [24] shadowed 8 users with visual disabilities for two days, and found similar concerns and workarounds. Some users were also concerned that their sensitive information being stored in an insecure manner which could leave them vulnerable to security breaches.

Assistive technologies affect users' experiences. For example, screen reader outputs are serial in nature; since information is delivered line-by-line, users with visual disabilities must sequentially listen to options to identify the desired item or must skip through headings and sample paragraphs until they have found relevant data to achieve their goals [45, 47].

When it comes to security, users with visual disabilities may not rely on HTTPS or SSL/TLS dialogues to assess whether a website is legitimate or fraudulent in Abdolrahmani et al.'s [1] study with 11 participants. Several expert evaluations have found that the security mechanisms involved in completing common web-based security tasks (like logging into a website or purchasing an item online) were inaccessible, impeded the opportunities for users with visual disabilities to behave securely, and could instill a false sense of security [14, 19, 33].

As a result, the security techniques of users with visual disabilities are different from sighted users' behaviours [45]. Most recent work in this realm has focused on novel security technology for users with visual disabilities. Voice-controlled assistants like Amazon Echo have become inadvertent accessible solutions for people with visual disabilities in independently managing smart devices in their homes and pose as aids for therapy and caregivers [30, 38]. Branham et al. [10] propose a number of design guidelines to adapt home assistants so that they are more efficient and controllable for users with visual disabilities. However, as Akter et al. [4] argue, smart home devices do not yet properly consider the contexts of assisting individuals with visual disabilities and should better consider the privacy and security of the users with visual disabilities and those in their environments.

Other recent security and privacy solutions have been more deliberately designed to aid people with visual disabilities, including: improved audio CAPTCHA implementations [20], observation-resistant password schemes [13, 31], and accessible password managers [7]. These technologies are successful because they leverage the unique capabilities of users with visual disabilities within the system design [43].

As noted in previous work [22, 48], this research area requires further investigation. Most studies in the area are conducted with small samples, which suggest that further validation is required. Additionally, many mainstream security and privacy mechanisms are still not designed to properly integrate the competencies of users with visual disabilities [43].

Our contributions to the literature: In this paper, we confirm and extend previous findings relating to the security concerns of users with visual disabilities. We further explore how they manage and interact with various security indicators on the web, and whether these actions offer the desired level of protection. We identify obstacles not yet discussed in the literature, including: assistive technology misleading users while they assess phishing indicators and an evident distrust by users in security advice. We discuss the complex nature of users' security management techniques and various factors contributing to risky behaviours sometimes forced by inaccessible indicators. Finally, we suggest recommendations for improving security mechanisms based on our research.

3 Methodology

Our study took place in 2018 and was cleared by our university's Research Ethics Board and the Canadian National Institute for the Blind (CNIB). Sessions took place in three quiet locations, with participants choosing the location most convenient: our research lab, conference room at the Canadian Council of the Blind (CCB), or an office at the CNIB. All sessions were audio-recorded and transcribed.

Phase 1: Pre-test Participants verbally completed a demographics questionnaire with the researcher.

Phase 2: Website Tasks Participants were asked to complete three security tasks on their assigned website (Table 1). If time permitted, participants were asked to repeat the process on a second website. Participants were pseudo-randomly assigned to websites ensuring even allocation across the three sites. Our protocol was that if a participant deemed a website illegitimate at any point, we told them to stop interacting with it and we assigned a new site. Participants worked on a task until they decided it was complete. Between each task, participants answered 5-point Likert scale questions about the usability of the task. The researcher noted any observations relating to participants' interactions with the websites.

Websites: The websites elicited opportunities for exposure to security risks pertaining to eCommerce and email. While the spoofed CNIB website is primarily an informational resource, the Shop and Donation pages collect personal information (e.g., address, credit card) which can put users' privacy at risk. The spoofed website used a domain we purchased, *ccnib.ca* and did not use SSL/TLS. Other than these differences, the spoofed website was identical to the legitimate one, in terms of the content and user interface design.

Technological setup: The technology used during the study varied according to participants' needs/preferences. We offered them two setups: a desktop with JAWS, ZoomText, keyboard, mouse, and speakers or, an iPad with built-in accessibility features. They could use these, plus any other tools (e.g., physical magnifying glass), or their own devices. One participant chose to complete the study on their own iPad, 12 used the desktop setup running Windows 10, and two used the iPad (iOS 11).

Collection of personal information: Personal information used in the tasks (e.g., (usernames, passwords, credit card information) was provided by the researcher. Participants were encouraged to complete the tasks as they normally would with their own information outside of the study. We avoided emphasizing security or privacy during the study, to mitigate bias on users' typical behaviours while interacting with the websites.

Phase 3: Questionnaires and semi-structured interviews Through two verbal questionnaires and a semi-structured interview, participants elaborated on their online security and privacy concerns, the security advice they have received, and the protective security actions they take in their everyday life outside of the study.

Questionnaires: The first questionnaire asked participants to rate (on a scale 1 to 5) their level of concern for each item in a list of cybersecurity threats mentioned in a previous study by other web users with visual disabilities [25], presented in random order. The second questionnaire asked participants to rate the effectiveness of common security advice [26, 40], and likelihood they would adhere to these protective actions in real life.

Interview: Next, we conducted a semi-structured interview to further investigate participants' most pressing concerns, methods for protecting themselves online, obstacles they face while maintaining their security and privacy. After the interview, we debriefed the participants who used the spoofed website.

3.1 Participants

Fourteen participants with visual disabilities (7 blind, 7 partially sighted), completed three phases of the 90-minute study. Participants were recruited via social media posts, mailing lists, and through the CNIB. Once recruited, participants were provided a digital copy of the consent form ahead of their session. At the beginning of the session, the researcher reviewed materials with participants, and obtained verbal consent.

Our participants (6 women, 8 men) were over 18 years old, from Ottawa or Toronto, and had a visual disability. Their median age was 52.5 years, similar to the age distributions of prior accessibility user studies [19, 42, 46]. Nine had a college diploma or university degree. We categorized eight as unemployed: they were full-time volunteers, on long-term disability, or active job seekers; six were employed.

Participants rated their limitations in three visual capability dimensions (see Table 3 in the Appendix): visual acuity, visual field, and light perception. Aligning with common usage of the terms, we categorized participants with "very limited" capabilities affecting both eyes as "blind" and others as "partially sighted." Participants were given $50 for their time and were compensated for study-related travel expenses.

All participants were familiar with using the Internet. In daily life, most blind participants relied on screen reading software such as JAWS, NVDA, and iOS VoiceOver. Those with partial vision used custom settings on their device/browser or used screen reading/magnifying software like ZoomText. Table 3 (Appendix) provides participant's demographics, and the technological setup they used during the study. Five with low vision used the ZoomText 11 screen magnifier; participants with low vision used no specialized assistive software and instead used custom browser settings or device features like pinch-to-zoom when needed. All blind participants used screen readers, like JAWS 18 or VoiceOver.

4 Results

We were able to holistically consider participants' experiences by gathering information from task-based observation,

Website	URL	Task A	Task B	Task C
Amazon	https://www.amazon.ca	Verify whether site is legitimate	Login (if safe)	Complete purchase
Gmail	https://mail.google.com	Verify whether site is legitimate	Login (if safe)	Download attachment
Spoofed CNIB	http://www.ccnib.ca	Verify whether site is legitimate	Find donation page	Donate money

Table 1: The websites used and associated tasks completed during the sessions. Note the extra C in the spoofed CNIB URL

questionnaires, and semi-structured interviews about their real life practices. We identified several usability and accessibility issues which impact users' capabilities to identify security threats and to employ protective actions.

We report on our findings from Phase 2 and 3 below, then we summarize the relationships identified between the various data into four states of online security and privacy awareness. These states depict general behavioural trends in our participants' experiences and touch upon the security and privacy threat scenarios related to these trends.

4.1 Phase 2: Website Tasks

Table 2 summarizes participants' accuracy in identifying the legitimacy of the websites and their self-reported responses for all website tasks during the study. These responses include: the perceived accessibility of the website, task ease, and confidence ratings. Confidence ratings related to participants' certainty in having completed the task in its intended entirety (i.e., correctly). We note that these represent participants' perspectives and do not necessarily reflect whether the task was actually completed successfully or securely.

Due to our sample size, we did not run statistical tests on website task data. However, generally, participants rated websites as accessible ($M = 4.0$, $SD = 1.3$), were confident they had completed the tasks correctly ($M = 4.5$, $SD = 0.8$), and thought that the tasks were neutral-to-easy to complete ($M = 3.9$, $SD = 1.1$). We focus on the obstacles observed.

Task A: The participants' first task was to verify the site's legitimacy. The Gmail and Amazon sites were legitimate, and the CCNIB site was a spoof. All but one participant considered the provided websites to be legitimate[1]; all reported a high degree of confidence in their assessments. As a result, participants were mostly correct about the legitimacy of Amazon and Google websites. However, *none* of the participants recognized our spoof website, CCNIB, as illegitimate. Overall, 11/18 assessments were correct despite participants' confidence in their ability to complete the task. In particular, all participants assessing CCNIB rated their confidence as 5 (on a 5-point Likert scale) despite their incorrect assessments.

We observed participants' legitimacy assessments to be impacted by several factors. First, many leveraged untrustworthy security indicators such as professional looking, or familiar sounding alternate text for, logos and page content. Secondly, some participants mentioned that the site seemed to be associated with a reputable organization so it must be legitimate and trustworthy. Thirdly, some participants were unsure how to assess website legitimacy because it was not something they often considered:

> *"I don't think about a site's security often. I would if it seemed like a hacky site. If it wasn't professional, or if things were out of order, or if the buttons were in weird places." (U03)*

Our observations suggest that many participants did not rely on trustworthy indicators when assessing website legitimacy.

Some participants did attempt to take security precautions that were aligned with security best practices. Specifically, we observed some participants double check URL addresses for spelling inconsistencies. Unfortunately and importantly, for blind participants, this effort was futile for the spoofed site since JAWS announced the spoofed CCNIB site's address in the same way it would read the legitimate CNIB URL: `H-T-T-P-colon-slash-slash-W-W-W-dot-cuh-nib-dot-cah`. Thus, blind participants could not detect this phishing clue unless they used the screen reader to read the URL letter by letter. Since it is unlikely for any user to do this unless they are already suspicious of a website, relying on JAWS feedback to detect domain inconsistencies is ineffective.

One partially sighted participant, U12, detected our phishing site's extra "c" by looking at the URL. However, they dismissed this concern and completed a monetary transaction because the page content met their expectations:

> *"There's a lot of detail here... I'm very confident that it is legitimate because I'm looking at a product [in their online store] that I'm familiar with, and that is really only sold by the CNIB." (U12)*

We were careful in our instructions to avoid priming participants to be unrealistically security-conscious, however, being part of the study may have encouraged some participants to let their guard down or otherwise behave in ways they would not outside of the laboratory setting in Phase 2. When asked to reflect on their behaviour related to the tasks, nearly all participants said it was similar to their real-life behaviour. Only one participant mentioned that that they trusted that the researchers would not "lead them astray" and were inclined to assume all provided websites were legitimate. While we took efforts to increase ecological validity and we have no

[1] The participant decided that the Gmail site was likely illegitimate only after completing all tasks.

indication that this was a widespread problem, this effect is a known challenge for security and privacy studies [23]. To accommodate for these limitations, we dive deeper into users' real life practices and attitudes during Phase 3.

Task B: All attempts (18/18) to complete Task B: logging in to Amazon and Gmail or finding the donation page on the CCNIB website were ultimately successful with some issues.

Participants experienced no issues with finding the donation page on CCNIB. The Gmail login page has minimal content and users are automatically placed in the login form fields. On the Amazon homepage, users must skim through page content to find the login link and then skim through the page to find the form fields for entering login credentials.

Despite their eventual success, blind participants experienced accessibility issues during the login processes with Gmail and Amazon. With JAWS' password masking techniques, each password character is announced as "star." This provided blind participants with no feedback to confirm which characters they had input. The websites also provided no audible feedback about successful login. Instead, participants relied on the lack of warning to confirm successful login. Some were initially unsure if they had successfully logged in the websites, or if they just could not find a warning about login failure when skimming elements from the entire page.

Additionally, participants were unsure how much sensitive information was being displayed on the screen after logging into the sites. This limited blind participants' control over account information and their personally identifiable information (PII) because they must audibly skim through the page to confirm successful login and may unintentionally instruct the screen reader to announce private account information aloud.

Task C: All participants were able to complete Task C on the Gmail and CCNIB websites but only four out of six participants were able to finalize a purchase through Amazon, giving an overall completion rate of 16/18.

The two participants who were unable to complete the task were blind: U04 used VoiceOver and U14 used JAWS. Both faced insurmountable accessibility obstacles on Amazon because of information provided only through colour-based cues. Specifically, the website formatted a corrected shipping address when finalizing a purchase. The nuanced differences between the original and corrected address were highlighted in red but not described with alternate text.

To progress through the purchasing process, users must choose one of the two formatted addresses. Both participants tried unsuccessfully to identify which address to use for several minutes before we guided them to the next portion of the study to ensure the remainder of the session could be completed within the study's allotted time.

This accessibility hurdle is another example of the limited control users with visual disabilities have over websites and, in turn, limited control over their PII while interacting with websites. In this circumstance, participants were unable to access feedback relating to issues with a mailing address. A blind user unable to perceive Amazon's suggested options could be forced to complete a task in a way they cannot be sure aligns with their security and privacy values (e.g., by sharing the task and access to their account with a sighted person for assistance). This can be concerning because users are often expected to understand the implications of their actions and may not be provided secure or private defaults [32].

4.2 Phase 3: Questionnaires

Phase 3 deals with participants' real life experiences, concerns, and attitudes. Figure 1 summarizes participants' reported level of concern for 12 cybersecurity threats common to people with visual disabilities. The number in each cell of the matrix indicates the number of participants who selected the given Likert scale response. The colour intensity of the cells is based on the popularity of the response, with higher numbers having darker colour intensity.

Our participants generally expressed moderately high levels of concern. They were most concerned with protecting their financial information, their identity, their data, and their device from theft or disclosure. They were least concerned with threats relating to surveillance and eavesdropping.

Figure 1 summarizes participants' Likert-scale responses for their perception of the effectiveness of each protective action and the likelihood that they would take these actions.

Participants rated most of the actions as effective or very effective for protecting themselves online. However, they gave low ratings to two fundamental security measures: enabling automatic updates, and using a password manager.

For both of these measures, participants identified accessibility issues that rendered them ineffective from their perspectives. For example, automatic updates can lead to system changes that cause programs to no longer be compatible with assistive software. Also, due to password masking techniques, JAWS announces password characters as a "star" rather than the character. This leaves blind participants unable to confirm the accuracy of their entered passwords before logging in or storing their passwords in software, which undermines the perceived utility of password managers. Allowing users to audibly unmask their typing when entering a password for storage into a password manager (or when logging in from a location safe from eavesdropping) might help with this issue.

We saw some relationships between participants' perceived effectiveness of advice and the likelihood that they would follow this advice: the actions rated as most effective were generally likely to be followed. However, this relationship was not true for all actions. Accessibility concerns had a direct impact on participants' likelihood to follow the protective actions. Participants were less likely to adhere to security advice they considered ill-fitting for people with visual disabilities. For example, while multi-factor authentication was considered

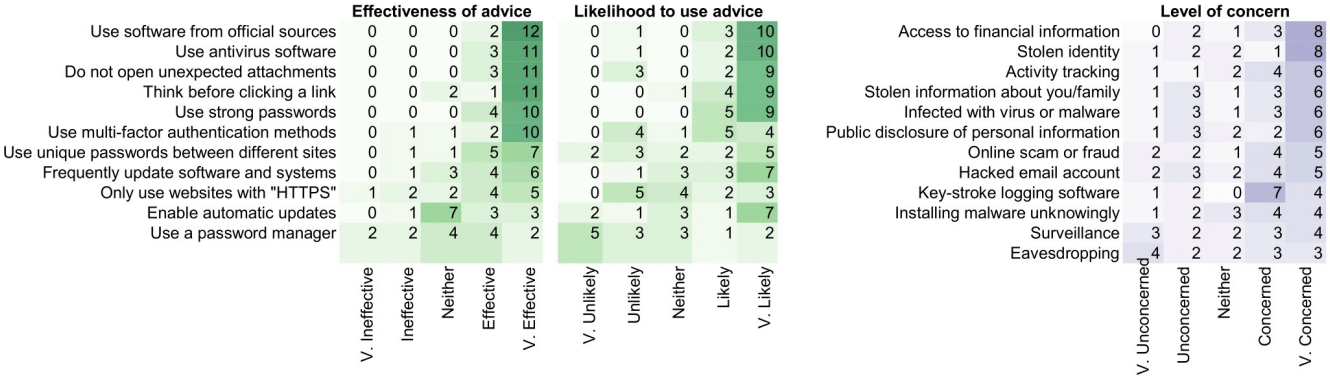


Figure 1: Number of participants selecting each Likert-scale response rating: the perceived effectiveness of security advice (left), likelihood that they will adhere to the advice (center), and their level of concern per threat (right). Darker cells indicate more popular responses.

very effective, many participants were unlikely to activate it on their own accounts because it increased the difficulty of logging in, and this task was already challenging on its own.

Some participants reported a lack of confidence in the security advice they receive. Two participants specifically noted their distrust in sighted individuals who present themselves as technology or web security experts. Participants expressed low confidence in the effectiveness of protective security actions and in the advice intended to help them avoid threats. Evident distrust in security advice was mainly rooted in a disconnect between the security expert's perception of the participant's experiences and participant's actual lived-experience:

> *"People say they know the difference between a threat and a non-threat, but someone who is actually blind knows the risk... People who use just regular everyday technology they take a lot of risks, it's just a reality. I have to be safer and smarter about it." (U01)*

Participants who expressed trust in security advice and tried to comply were greatly hindered by accessibility issues. For example, U14 explained that he used anti-virus software and kept the program updated. Yet, aspects of the interface were inaccessible to his screen reading software so he was unable to read and resolve flagged issues. The participant expressed that when confronted with a warning, he had to chose from the subset of accessible actions within his antivirus and hope that these would resolve the detected issue.

In some cases, adhering to security advice is not an option. When U12, a partially sighted participant, attempted to input information on the CCNIB website, he was unable to easily locate the form fields because the website was incompatible with the Chrome plugins he used to increase page contrast and aid in identifying page sections.[2] U12 explained that this was a common issue that often forced him to move closer to the screen and strain his only sighted eye. In these circumstances, he could not prioritize protecting sensitive data:

> *"I'm literally just hoping this will work. So safety's not really being considered, which is unfortunate, obviously." (U12)*

These findings demonstrate a need for improved mechanisms and security advice which properly consider the circumstances of users with visual disabilities and the assistive software they use. Improvements may increase users with visual disabilities' trust in the system, and enable them to perform the security actions that they wish to undertake.

4.3 Phase 3: Interviews

We further collected contextual data relating to users' real life experiences while browsing online, the strategies they employ to maintain their security online, and their feelings of safety.

4.3.1 Qualitative Analysis Methodology

Interview data, observational notes, and other verbal feedback provided during the session was coded based on Braun and Clarke's [11] six phases of thematic analysis.

Our initial research intent was to explore themes relating to users' attitudes, behaviours, concerns, and desires. For the first iteration, the lead researcher extracted 356 relevant excerpts from notes and recordings from all participants. These were organized according to the four overarching themes with closely related excerpts grouped as trends. We coded trends to formulate an initial codebook containing 35 codes. Three researchers iteratively discussed and refined the codebook, resulting in 17 codes in the second version.

[2] Note that we had duplicated the legitimate site exactly, and that the CNIB site **should** be accessible given that its target users have visual disabilities.

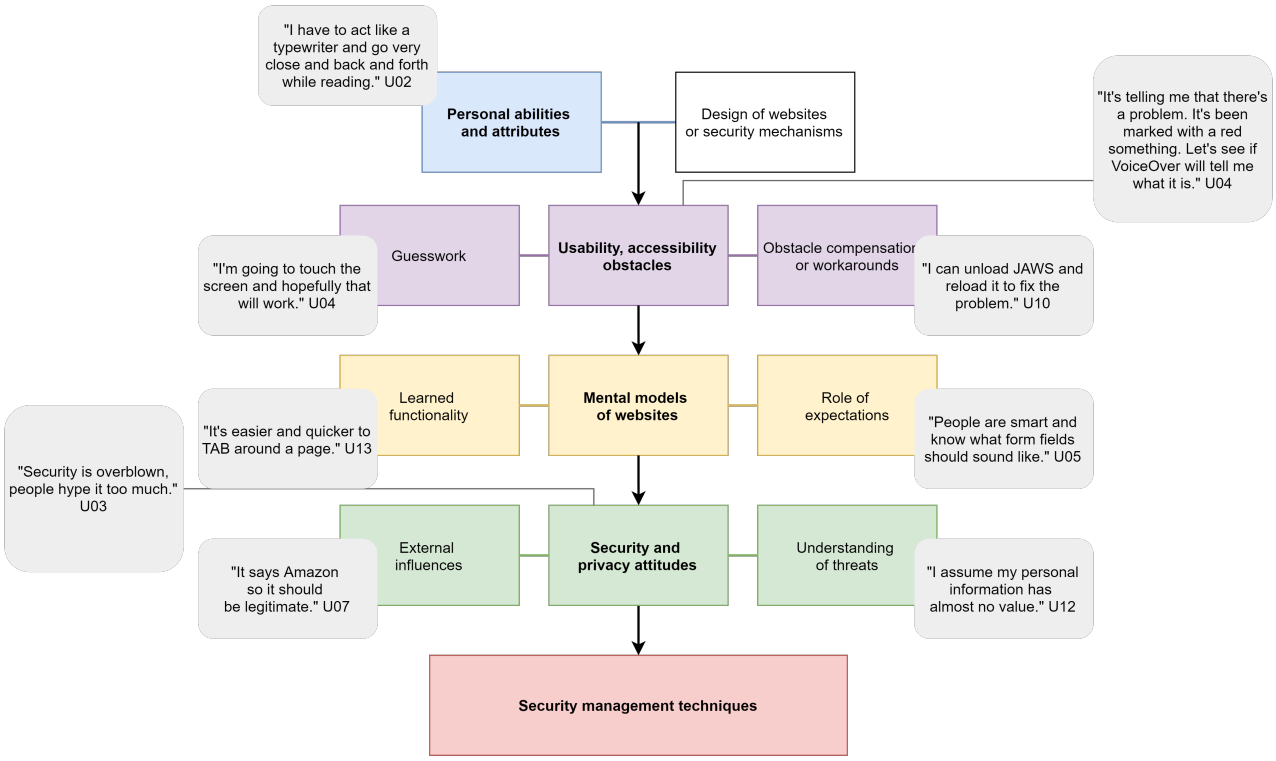


Figure 2: Relationships between the main codes formulated during our thematic analysis of participant interviews and feedback comments. Example excerpts are also included.

During the second round, two of these researchers used the second version of the codebook to independently code the same five randomly selected 90-minute transcripts. The mean Kappa score for inter-coder agreement across all codes was 0.66. This can be interpreted as good agreement. We met to discuss discrepancies in coding, come to agreement, combine redundant codes, and modify others to better fit the data, resulting minor changes. Then, the remaining transcripts were split between the two researchers and coded with the final codebook as shown in Table 4 in the Appendix.

4.3.2 Results

We summarize key takeaways from our qualitative analysis in Figure 2. Our analysis suggests an interdependent nature amongst our codes which linked to participants' security management techniques (further explored in Section 5).

Below, we provide sample excerpts to describe key codes/code groups and their relationships. In examining the relationships, we noted similarities with Cranor's human-in-the-loop security framework [18] which details aspects of effective security communications and can be used to identify how security indicators may fail to deliver information to users. Notably, we identified parallels with the *personal variables, intentions,* and *capabilities* factors which affect how users receive, process, and apply security and privacy information. When framed within this context, our qualitative analysis can provide insight into the nuances of communicating security information to users with visual disabilities.

Personal abilities and attributes: During discussion, participants contextualized the obstacles they faced with details about their visual disabilities, preferences, personalities, or technological skills when handling obstacles:

> *"As a partially sighted person, I try to get rid of the clutter, even in my mind, before I do something like this because it's easy to get distracted and take more time or more unnecessary use of vision." (U12)*

> *"I guess I should be a little more vigilant but, I'm still one of those people that if I go on a website I assume that that's where I should be." (U02)*

> *"I'm used to problem solving text stuff. That's what I do, and that's what I teach other people to do so it doesn't bother me that much. I just wish I could do it faster." (U05)*

Demographics and personal characteristics impact a person's ability to understand security indicators and influence how they take protective actions [18]. Participants' feedback in this code group was critical to understanding the context of users' experiences and fundamental variables impacting other codes relating to users' security/privacy attitudes and

behaviours. Thus, Figure 2 has it as the highest level factor in the chain influencing security management techniques.

Usability, accessibility obstacles: Participants described several challenges they experienced which were influenced by their individual capabilities and characteristics. We recognized these issues as infringements of basic usability or accessibility principles. These issues, when framed as *communication impediments* [18] can cause partial or full security information communication failures.

We gathered further insight relating to these obstacles as participants speculated what went wrong and described the workarounds they used to achieve their goals:

> *"I guess the webpage was programmed such that this was worthy of a restart... I don't think it had to do with something we tapped on. I think it had to do with the way the page was structured." (U04)*

> *"I can unload JAWS and reload it because that will fix the problem. If it doesn't read anything like before, I restart it again." (U10)*

As shown in Figure 2, the obstacles users faced were shaped by their individual characteristics and then influenced how they perceived and operated websites. For example, blind participant with technical backgrounds who faced several accessibility issues described more sophisticated workarounds, such as using advanced search options, compared to partially sighted users who encountered fewer issues. Users with sophisticated workarounds also mentioned using technical security indicators such as HTTPS or checking for SSL/TLS certificates. Interestingly, sophisticated workarounds did not necessarily align with accurate interpretations of how these indicators help their security, suggesting that these users had a superficial grasp of the issue despite their background.

Mental models of websites: The literature suggests that a user's familiarity with security indicators, vocabulary, and structure will impact their comprehension of risks and threats [18]. Thus, participants' feedback relating to how they understand and use systems was essential.

Participants described the shortcuts they use to interact with websites and browsers such as skimming for relevant page content via headings and using tabs to quickly access page features. Our findings reflected the shortcuts noted in related studies [19, 43] exploring the browsing behaviours of users with visual disabilities. These excerpts also reflected the importance of consistency and standard presentation as new interfaces can take a long time to learn:

> *"I assume that the actions are going to be on the right-hand side of the margin. If I was clueless and I didn't know how to use a website at all, then I would be bouncing around there for days." (U01)*

As highlighted in Figure 2 participants' mental models were impacted by the obstacles they faced and their mental models subsequently had downstream implications on how they completed security tasks. At times, participants had developed useful heuristics to inform their mental models and potentially help them with identifying phishing:

> *"You can usually tell if something you're looking at [isn't] actually Google or PayPal because it will say your bank account is compromised, click here. Banks never do that. They won't say click here to go to your account." (U05)*

Other times, participants' mental models and expectations for websites included reliance on unreliable cues that could mislead them. This also occurs with sighted users, but the types of cues occasionally differed because of the lack of visual feedback. For example, a blind participant believed a website was legitimate when they heard form feedback they had previously heard while using another website they trusted. Similarly, two partially-sighted participants trusted the spoofed CNIB website after they found content about assistive technology for people with visual disabilities, and this aligned with their expectations for the website.

Security and privacy attitudes: We coded participants' relevant comments while completing website tasks and questionnaires relating to their security concerns and advice. During the interviews, participants provided further information about what made them feel secure while browsing online. These included external influences like trusting specific companies or trusting friends and family:

> *"I feel safe online when people I've trusted tell me that whatever I'm using is safe. Anti-virus will keep me safe. My passwords will keep me safe. Sticking to what I know will keep me safe." (U01)*

> *"I feel safe on pages [where] I'm offering sensitive information, I believe that a company will have something to lose. If I lose, they lose too." (U04)*

Participants' security and privacy attitudes were also impacted by their understanding of technology and by their understanding of associated security threats:

> *"On my phone I know I'm not going to get viruses. I open attachments on my phone so I can save it to Dropbox or somewhere where I can access it on any platform. That way I'm not getting viruses or anything I don't need to have." (U05)*

Participants rated the threats they found most concerning in Section 4.2, and we found broad agreement for some threats. When we probed this topic further during the interviews, we noted that, despite some agreement on the scales, some participants explained that they found these issues very concerning, whereas others expressed an unconcerned attitude towards security and privacy:

"I always have to have my guard up. I know that people would perceive me as vulnerable." (U10)

"I don't really think about security because if I would always think about the, 'Oh what would happen if...', then I would never go online." (U07)

Attitudes seemed to be influenced by participants' individual characteristics, the obstacles they faced, and their individual security mental models. The different attitudes also contributed to the differences we observed in users' **security management techniques** which we elaborate in Section 5.

5 States of Online Security and Privacy

Security and privacy management techniques can be viewed as an amalgamation of users' lived experiences and understanding of websites/security mechanisms which are limited by accessibility and usability obstacles. Participants' adapt their security and privacy strategies depending on several factors relating to personal experiences and external factors. An individual may transition between strategies depending on the context of the task at hand, or may get stuck in one state due to accessibility obstacles or their security mental models.

The relative importance of each factor in influencing security management techniques varied per participant. Individual participants' management techniques also changed depending on the accessibility issues they faced per website, their current task goals, and the value of the information they exchanged with websites. To address the fluidity and complexity of this process, we identify "states" of security and privacy awareness that participants may go through while browsing online and affect their related behaviours and strategies.

These states are relevant to participants with any degree of vision disability as we did not observe that this influenced their likelihood of being associated with a given state. Furthermore, similar to describing security folk models [49], we focus less on the accuracy of participants' perspectives and more on the potential security and privacy implications related to these states of awareness.

5.1 Unconcerned, overconfident

Participants in the *unconcerned, overconfident* state either believed that they had taken the necessary precautions and that they could now freely navigate online without risk, or they believed that it was easy to spot online risks so additional precautions were unnecessary. In both cases, participants were unknowingly placing themselves at risk.

As previously mentioned, participants' understandings of security was greatly influenced by their understanding of web technology and the security mechanisms enabled on their system. Specifically, U04 shared that after taking precautions to protect himself and his devices, he is not concerned about his security and privacy and thus proceeds to trust that he will be secure while completing tasks online. However, some precautions U04 implemented relate to a common misconception that Apple products are impervious to security breaches.

Other participants made similar comments relating to Apple products or websites affiliated with Amazon or Google. Additionally, those who expressed lower levels of concern tended to rely on gut reactions about which websites seemed "hacky" and unprofessional when detecting threats. These assessments rely solely on website content they can read with assistive technology and cannot include available information that may be helpful but is inaccessible. This suggests that individuals relating to this state of security and privacy awareness may be more likely to fall victim to social engineering techniques relying on high-fidelity copies of the legitimate site, or spoofed organizational affiliations while completing tasks online. Therefore, an unconcerned, overconfident approach to security can lead to increased risk-taking habits or, in the worst case scenario, security apathy such as the following:
"Security is overblown. People hype it too much." (U03)

5.2 Concerned, overwhelmed

Participants in the *concerned, overwhelmed* state were worried about their online security and privacy but were unsure which protective techniques could address their concerns and were not confident in their ability to protect themselves online.

These participants expressed deep concern regarding their online security and privacy. Individuals relating to this state were more likely to mention security and privacy considerations while completing tasks. Additionally, these individuals mentioned several repressive habits they have in real life, including not banking or shopping online, only visiting websites which were recommended by trusted family or friends, and deleting all emails received from unknown recipients because they did not trust their own abilities in detecting threats. Often, individuals who relate to this state were anxious because of personal or secondhand experiences with security breaches.

Individuals who demonstrated great concern regarding their security and privacy also often expressed uncertainty in their ability to identify potential threats and to implement effective protections due to conflicting advice or accessibility issues that hindered them from taking desired precautions. Once in this state, an individual may feel overwhelmed or blame themselves for this uncertainty:

"Because I don't have any kind of background in programming or anything other than just being an end-user, I feel like a lamb to the slaughter. I just go in there without knowing that I shouldn't be." (U02)

5.3 Jaded, resigned

Participants in the *jaded, resigned* state may have been concerned about their online security and privacy, but severe us-

ability issues forced them to abandon protective actions and rely on others to protect their online security and privacy.

These participants approached their online security and privacy with fatigue due to usability and accessibility issues which limited their ability to employ protective strategies. These participants expressed a sense of powerlessness and were ultimately forced to rely on other, sighted, individuals to manage their security and privacy. Particularly, U14 regularly faced accessibility obstacles in managing his antivirus software, updating his systems, and navigating websites. Ultimately, he relied on his daughter to verify his security when completing tasks. Similarly, when U09 faces major challenges, she must relinquish autonomy and rely on trusted family members and friends to complete online purchases on her behalf to assure the security of her financial information.

Those in a jaded state initially approach their online activities with concern and try to be proactive against threats, but may become resigned:

> *"If someone wants to hack your computer, they will do it because there are always loopholes in any software that you're using. It doesn't matter whether you have the best antivirus or security software, it can still be hacked." (U13)*

When individuals are concerned for their security and privacy but must forfeit their independence to complete tasks, their ability to engage with technology is greatly limited.

5.4 Comfortable, unimpeded

Some participants were confident in the actions they took to protect themselves online while others were less inhibited by accessibility obstacles. Yet, no participants were both completely unimpeded, comfortable, and used effective security management techniques. Therefore, this fourth security state relates to an ideal state wherein users with visual disabilities are technologically empowered and can confidently manage their online security and privacy.

Users relating to this security state would have readily available access to all pertinent information they need to form informed security and privacy decisions. Additionally, users with visual disabilities in this state would have access to advice about protecting their security and privacy which adequately considers their nuanced concerns and lived realities relating to non-visual browsing experiences and the interaction between websites/software and assistive technologies. Furthermore, individuals in this state would be familiar with protective best practices and be able to implement these tactics in a manner that better reflects their browsing strategies.

To reach this state, we need to better consider the unique strengths and capabilities of different user groups, including those with visual disabilities in the design of security and privacy interfaces. Aligning with Reyez-Cruz et al. [43], we suggest that more sophisticated designs should include modes of interaction ideally suited to the capabilities of different groups of users rather than simply considering accessibility as an add-on to the "standard" interface.

5.5 Comparing to Sighted Users

We briefly highlight the main commonalities between our findings and related literature on sighted users. For example, *optimism bias and overconfidence* [2] refers to users underestimating the chances of becoming a victim to cybercrime and thus becoming less alert online. Like users with visual disabilities in the *Unconcerned, overconfident* state, sighted users who are familiar with a website may feel safe, trust that they are secure, and then bypass warnings [41]. This bias puts both groups of users at risk especially when they use unreliable cues like website content [5] to decide whether a website is legitimate. However, we note that sighted users may have more opportunity to recognize and recover from their error since most security cues are visual.

Furthermore, users may not have not enough mental resources to evaluate all options and potential consequences while attempting to achieve their goals [2] and must sift through overwhelming amounts of advice to make security and privacy decisions [39]. While these studies were done with sighted users, we note some parallels with participants from our study falling into the *Concerned, overwhelmed* state. Again, the differentiating factor is the additional burden faced by users with visual disabilities who must also deal with accessibility challenges and security advice that makes assumptions about users' ability to view security cues.

6 Discussion

Through task-based scenarios, questionnaires, and semi-structured interviews, we uncovered several major usability issues for users with visual disabilities. Users were hindered from completing security activities during the study and in real life, including accurately verifying the legitimacy of a website, securely logging into a website, and maintaining control of PII while completing online transactions. These obstacles impeded users' mental models of websites and negatively impacted their security and privacy attitudes.

Our study focused on strengthening the empirical knowledge base of accessibility issues pertinent to online security tasks. Particularly, we confirm findings relating to the security and privacy concerns of users with visual disabilities [3], their website credibility assessments [1], and the role of sighted allies when managing online security and privacy [24]. Our work increases confidence in the generalization of research findings within the realm of usable security and accessibility. This triangulation and confirmation work is particularly important given that studies in this area often have small sample sizes due to the difficulties of recruiting for this population.

Our study also extends existing work by highlighting several instances where security information is not effectively

communicated to users via assistive technology. Furthermore, participants identified ill-fitting security advice they perceived as ineffective and were unlikely to employ. Participants also shared their experiences with inaccessible indicators and anti-virus software. Further, we observed that interfaces provided participants with little to no guidance for protecting themselves online and, at times, they were prevented from completing their task entirely or were misled by assistive cues (e.g., reading the CCNIB URL in an identical manner as the legitimate CNIB URL). To our knowledge, the observation that assistive technology can actually mislead users or obfuscate important security cues has not previously been reported.

Some of the issues raised in our study could be avoided by adhering to website accessibility guidelines, but we note that the issue is more complex than this. These guidelines do not address the unique issues that arise in supporting users while maintaining their online security and privacy. One significant factor is that online security relies on more than the design of a website itself, which is the sole focus of most guidelines. For example, accessible web security also involves the browser chrome and other software or mechanisms (e.g., antivirus software, password manager), as well as the interaction between these technologies and the assistive software.

We outline recommendations for designing security interactions which can better serve users with visual disabilities in transitioning towards more beneficial states of privacy and security awareness. These recommendations align with existing general guidelines in usable security, and focus on the nuances of applying these principles when considering users with visual disability. We also emphasize the importance of closely collaborating with people with visual disabilities, ideally who are knowledgeable about security, to ensure that any changes resulting from these recommendations properly reflect the perspective and needs of users with visual disabilities.

Prioritize security information: Security interfaces should describe the current state of security and related available functions in simple and clear language [50]. Much of this information is available in browsers but cannot not be accessed by users with visual disabilities due to a mismatch between the competencies of these users and the design of most security interfaces. Sometimes this information is overlooked by users while trying to compensate for other accessibility issues. Therefore, security information should be more readily available via different modalities in a prominent, predictable, and easy-to-access location.

Assistive software output could prioritize security information over page content. For blind users, this would mean that reliable indicators are read aloud before less reliable indicators like page titles or content. For partially sighted users, this information could be pushed into, and emphasized within, their default field of view such as automatically zooming in on an address bar or other visual security cues rather than the page's header or navigation menu. Designers could also take advantage of the sequential nature of the web page experiences of users with visual disabilities. If properly implemented, users would automatically scan through security indicators before accessing the web content. This will inform users of potential security measures (or risks) before they interact with page content and decide to trust a website. However, designs will have to carefully balance the priorities of users to avoid potential frustration caused by presenting security warnings before relevant task information. Ultimately, users should retain control over whether security information is prioritized or simply easily available on-demand.

The use of other sensory channels can be used to minimize competition between website content and security cues. Salient non-visual warnings, like temperature feedback [34,51] can aid users with and without visual disabilities.

Provide proactive assistance: Security systems should be designed in a way that users can diagnose and recover from security errors [16]. Our work shows that screen reader users were not provided sufficient audible information to properly diagnose errors that were visibly shown on the tested websites. Some mentioned being unable to access and comprehend the problems being flagged by their anti-virus software. All of our participants demonstrated a willingness to resolve issues, but were uncertain of how to properly recover from the errors they faced. We emphasize that cues which help users in fixing security issues should be both accessible and directive.

Directive systems should proactively suggest solutions to users while providing enough context that they can understand the current state of their system and, if needed, how to improve it, without negatively impeding their cognitive load. This suggestion is based on: (i) the evident mental models of our participants with visual disabilities, (ii) their expressed need for more helpful guidance, and (iii) Felt et al.'s "suggestive design" approach to SSL/TLS dialogues [21]. In the context of sighted users, Felt et al. argue that users are more likely to adhere to security warnings if the dialogues highlight the advised steps. Directive security and privacy mechanisms can help users with visual disabilities who are concerned but unsure how to protect themselves online. Improved guidance can prevent these individuals from transitioning to a state of feeling helpless and resigned.

Similar to prioritizing security information, proactive assistance has the potential to cause frustration if delivered at an inopportune time. Users who are already at capacity with their current task may be overwhelmed by additional information, no matter how well intended. Making the assistance available in a side channel accessible on-demand may be preferable to interrupting the user's primary task. Future work should explore how to best assist users with visual disabilities who desire further support.

Make security advice relevant: Many of our participants with visual disabilities completed online transactions with an

inherent trust in their devices and/or the organizations that supposedly owned the websites. Sighted users also trust that external entities (E.g., firewalls, IT staff, or website owners) will maintain proper security [12,49]. Due to the severe accessibility obstacles, users with visual disabilities currently have limited means to personally maintain their security. Thus, interfaces which provide accessible contextual security and privacy guidance could be helpful for these users.

Security advice for users with visual disabilities must appropriately fit their lived experiences. Users who did not perceive sources of advice to empathize with their experiences and circumstances were unlikely to employ suggested security best practices. Future work could develop better security advice tailored for people with visual disabilities and the realities of their online experiences and assistive software. Participatory techniques are necessary wherein individuals with visual disabilities collaborate with sighted counterparts to devise appropriate tools and materials [43].

6.1 Recommended Practices

Reflecting on our practices, we identify some aspects of our study that facilitated participation for our target user group. In particular: recruiting through a trusted advocacy group, having the option to meet participants at a familiar place, covering the cost of transportation for the participant and an aid if necessary, providing the option to use their own devices, allowing individuals to self-identify whether they met the study's participation criteria, and avoiding unnecessary stress and risk from using their personal credentials (which could be visible to the researchers).

For this study, we worked closely with CNIB while designing, recruiting, and facilitating our study. We emphasize that the perspective of individuals within the target community should heavily influence all aspects of the research. Ideally, these individuals should be members of the research team. When not feasible, working closely with an advocacy organization like the CNIB, or community groups (e.g., Hayes et al. [24]), offers a viable alternative. We recommend that interested readers reference some of the excellent literature on conducting respectful and cooperative research involving people with have visual disabilities (e.g., [3,8,24,43]).

6.2 Limitations

Our findings provide insight to the behaviours and attitudes of users with visual disabilities. We collected data from a sample of local individuals whose views may not fully reflect the experiences of all people with visual disabilities. Additionally, our sample size is similar to those in related literature but is small compared to other usability studies due to recruitment difficulties despite our collaboration with CNIB. Furthermore, lab studies can introduce biases relating to users' behaviours or self-reported responses. Particularly, participants were provided credentials to complete tasks. This may have led participants to be less cautious; however, all participants said that they behaved in study as they normally would in real life. To further counter this potential bias, we focused a large part of our analysis on questionnaire and interview data exploring their real-life practices, in addition to observations from the the study tasks.

Future studies could explore alternative methodologies (e.g., using throw-away accounts or linking study compensation to performance) but these have their own trade-offs and limitations. Alternatively, studies could leverage other data collection methods, such as indirect observation [28] over a longer time period to further monitor how users behave outside of the lab. Additionally, accessibility and security research should go beyond considering visual disabilities to consider other disabilities and their intersections.

7 Conclusion

Through task-based scenarios, questionnaires, and a semi-structured interview with users who have visual disabilities, we identified a number of significant barriers they face while managing their online security and privacy, including: inaccessible antivirus software, misleading screen reader outputs, insufficient feedback relating to login processes, and unsuitable security advice. Participants' real life online security and privacy strategies varied depending on their current state of security and privacy awareness. Some people were prone to risk-taking habits and security apathy due to their trust in particular devices or associated organizations. Others were more concerned but felt unsure and overwhelmed while trying to protect themselves. Often, these individuals did not trust that they had the abilities to identify potential threats nor trust security advice that did not reflect their lived experiences. Obstacles led to security fatigue in some cases, where some users with visual disabilities felt resigned to rely on trusted sighted family and friends to manage their online interactions. Future work should continue to explore how to improve currently implemented security mechanisms with better consideration of a wider range of users' needs and capabilities.

Acknowledgments

The authors acknowledge funding from Natural Sciences and Engineering Research Council of Canada (NSERC) through the Canada Graduate Scholarships Doctoral program (Napoli), Discovery Grant program (Chiasson), and Canada Research Chair program (Chiasson). This research was also supported by an eCampusOntario Digital Inclusion Research Grant for 2017-18.

References

[1] A. Abdolrahmani and R. Kuber. Should I trust it when I cannot see it?: Credibility assessment for blind web users. In *ASSETS*, pages 191–199. ACM, 2016.

[2] A. Acquisti, I. Adjerid, R. Balebako, L. Brandimarte, L. F. Cranor, S. Komanduri, P. G. Leon, N. Sadeh, F. Schaub, M. Sleeper, Y. Wang, and S. Wilson. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Comput. Surv.*, 50(3), August 2017.

[3] T. Ahmed, R. Hoyle, K. Connelly, D. Crandall, and A. Kapadia. Privacy concerns and behaviors of people with visual impairments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 3523–3532. ACM, 2015.

[4] T. Akter, B. Dosono, T. Ahmed, A. Kapadia, and B. Semaan. "I am uncomfortable sharing what I can't see": Privacy concerns of the visually impaired with camera based assistive applications. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1929–1948. USENIX Association, August 2020.

[5] M. Alsharnouby, F. Alaca, and S. Chiasson. Why phishing still works: User strategies for combating phishing attacks. *International Journal of Human-Computer Studies*, 82:69–82, 2015.

[6] R. Babu, R. Singh, and J. Ganesh. Understanding blind users' web accessibility and usability problems. *AIS Transactions on Human-Computer Interaction*, 2(3):73–94, 2010.

[7] N. Barbosa, J. Hayes, and Y. Wang. Unipass: design and evaluation of a smart device-based password manager for visually impaired users. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 49–60. ACM, 2016.

[8] N. Barbosa and Y. Wang. Lessons learned from designing and evaluating smart device-based authentication for visually impaired users. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, Denver, CO, June 2016. USENIX Association.

[9] Y. Borodin, J. Bigham, G. Dausch, and I. Ramakrishnan. More than meets the eye: A survey of screen-reader browsing strategies. In *Proceedings of the 2010 International Cross Disciplinary Conference on Web Accessibility (W4A)*, page 13. ACM, 2010.

[10] S. M. Branham and A. Rishin M. Roy. Reading between the guidelines: How commercial voice assistant guidelines hinder accessibility for blind users. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 446–458, New York, NY, USA, 2019. Association for Computing Machinery.

[11] V. Braun and V. Clarke. Thematic analysis. In H. Cooper, P. Camic, D. Long, A. Panter, D. Rindskopf, and K. Sher, editors, *APA handbook of research methods in psychology*, volume 2, chapter 4. American Psychological Association, Washington, DC., 2012.

[12] C. Bravo-Lillo, L. F. Cranor, J. Downs, and S. Komanduri. Bridging the gap in computer security warnings: A mental model approach. *IEEE Security & Privacy*, 9(2):18–26, 2010.

[13] D. Briotto Faustino and A. Girouard. Bend passwords on bendypass: A user authentication method for people with vision impairment. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '18, page 435–437, New York, NY, USA, 2018. Association for Computing Machinery.

[14] M. Buzzi, M. Buzzi, B. Leporini, and F. Akhter. User trust in ecommerce services: perception via screen reader. In *New Trends in Information and Service Science, 2009. NISS'09. International Conference on*, pages 1166–1171. IEEE, 2009.

[15] Statistics Canada. Partcipation and activity limitation survey 2006 facts on seeing limitations. Technical report, Canada, 2006.

[16] S. Chiasson, P. van Oorschot, and R. Biddle. A usability study and critique of two password managers. In *USENIX Security Symposium*, pages 1–16, 2006.

[17] S. Chiasson, P. van Oorschot, and R. Biddle. Even experts deserve usable security: Design guidelines for security management systems. In *SOUPS Workshop on Usable IT Security Management (USM)*, pages 1–4. USENIX, 2007.

[18] L. F. Cranor. A framework for reasoning about the human in the loop. *UPSEC*, 8(2008):1–15, 2008.

[19] B. Dosono, J. Hayes, and Y. Wang. "I'm stuck!:" a contextual inquiry of people with visual impairments in authentication. In *Proceedings of The Symposium on Usable Privacy and Security*, pages 151–168. USENIX, 2015.

[20] V. Fanelle, S. Karimi, A. Shah, B. Subramanian, and S. Das. Blind and human: Exploring more usable audio CAPTCHA designs. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pages 111–125. USENIX Association, August 2020.

[21] A. Felt, A. Ainslie, R. Reeder, S. Consolvo, S. Thyagaraja, A. Bettes, H. Harris, and J. Grimes. Improving ssl warnings: Comprehension and adherence. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2893–2902. ACM, 2015.

[22] O. Gaggi, G. Quadrio, and A. Bujari. Accessibility for the visually impaired: State of the art and open issues. In *2019 16th IEEE Annual Consumer Communications Networking Conference (CCNC)*, pages 1–6, 2019.

[23] S. Garfinkel and H. Lipford. Usable security: History, themes, and challenges. *Synthesis Lectures on Information Security, Privacy, and Trust*, 5(2):1–124, 2014.

[24] J. Hayes, S. Kaushik, C. E. Price, and Y. Wang. Cooperative privacy and security: Learning from people with visual impairments and their allies. In *Symposium on Usable Privacy and Security ({SOUPS})*. USENIX, 2019.

[25] F. Inan, A. Namin, R. Pogrund, and K. Jones. Internet use and cybersecurity concerns of individuals with visual impairments. *Journal of Educational Technology & Society*, 19(1):28, 2016.

[26] I. Ion, R. Reeder, and S. Consolvo. "... no one can hack my mind": Comparing expert and non-expert security practices. In *Symposium On Usable Privacy and Security*, volume 15, pages 1–20. USENIX, 2015.

[27] P. Jaferian, D. Botta, F. Raja, K. Hawkey, and K. Beznosov. Guidelines for designing it security management tools. In *Proceedings of the 2nd ACM Symposium on Computer Human interaction For Management of information Technology*, page 7. ACM, 2008.

[28] T. Jagatic, N. Johnson, M. Jakobsson, and F. Menczer. Social phishing. *Communications of the ACM*, 50(10):94–100, 2007.

[29] J. Lazar, A. Allen, J. Kleinman, and C. Malarkey. What frustrates screen reader users on the web: A study of 100 blind users. *International Journal of Human-Computer Interaction*, 22(3):247–269, 2007.

[30] B. Leporini and M. Buzzi. Home automation for an independent living: Investigating the needs of visually impaired people. In *Proceedings of the Internet of Accessible Things*, W4A '18, New York, NY, USA, 2018. Association for Computing Machinery.

[31] D. Marques, T. Guerreiro, L. Duarte, and L. Carriço. Under the table: tap authentication for smartphones. In *Proceedings of the 27th International BCS Human Computer Interaction Conference*, page 33. British Computer Society, 2013.

[32] A. H. Mhaidli, Y. Zou, and F. Schaub. "We can't live without them!" app developers' adoption of ad networks and their considerations of consumer risks. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, Santa Clara, CA, August 2019. USENIX Association.

[33] D. Napoli. Developing accessible and usable security (ACCUS) heuristics. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, pages SRC16:1–SRC16:6, New York, NY, USA, 2018. ACM.

[34] D. Napoli, S. Navas Chaparro, S. Chiasson, and E. Stobert. Something doesn't feel right: Using thermal warnings to improve user security awareness. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. USENIX Association, August 2020.

[35] J. Nurse, S. Creese, M. Goldsmith, and K. Lamberts. Guidelines for usable cybersecurity: Past and present. In *Cyberspace Safety and Security (CSS), 2011 Third International Workshop on*, pages 21–26. IEEE, 2011.

[36] World Health Organization. World report on vision, 2019.

[37] C. Power, A. Freire, H. Petrie, and D. Swallow. Guidelines are only half of the story: accessibility problems encountered by blind users on the web. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 433–442. ACM, 2012.

[38] A. Pradhan, K. Mehta, and L. Findlater. "Accessibility came by accident": Use of voice-controlled intelligent personal assistants by people with disabilities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–13, New York, NY, USA, 2018. Association for Computing Machinery.

[39] E. M. Redmiles, S. Kross, and M. L. Mazurek. How i learned to be secure: A census-representative survey of security advice sources and behavior. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 666–677, New York, NY, USA, 2016. Association for Computing Machinery.

[40] R. Reeder, I. Ion, and S. Consolvo. 152 simple steps to stay safe online: Security advice for non-tech-savvy users. *IEEE Security & Privacy*, 15:55–64, 2017.

[41] R. W. Reeder, A. P. Felt, S. Consolvo, N. Malkin, C. Thompson, and S. Egelman. An experience sampling study of user reactions to browser warnings in the field. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page

1–13, New York, NY, USA, 2018. Association for Computing Machinery.

[42] G. Regal, E. Mattheiss, M. Busch, and M. Tscheligi. Insights into internet privacy for visually impaired and blind people. In *International Conference on Computers Helping People with Special Needs*, pages 231–238. Springer, 2016.

[43] G. Reyes-Cruz, J. E. Fischer, and S. Reeves. Reframing disability as competency: Unpacking everyday technology practices of people with visual impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery.

[44] D. Rømen and D. Svanæs. Validating WCAG versions 1.0 and 2.0 through usability testing with disabled users. *Universal Access in the Information Society*, 11(4):375–385, 2012.

[45] N. Sahib, A. Tombros, and T. Stockman. A comparative analysis of the information-seeking behavior of visually impaired and sighted searchers. *Journal of the Association for Information Science and Technology*, 63(2):377–391, 2012.

[46] S. Szpiro, S. Hashash, Y. Zhao, and S. Azenkot. How people with low vision access computing devices: Understanding challenges and opportunities. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 171–180. ACM, 2016.

[47] M. Vigo and S. Harper. Challenging information foraging theory: Screen reader users are not always driven by information scent. In *Conference on Hypertext and Social Media*, pages 60–68. ACM, 2013.

[48] Y. Wang. The third wave? Inclusive privacy and security. In *Proceedings of the 2017 New Security Paradigms Workshop*, NSPW 2017, page 122–130, New York, NY, USA, 2017. Association for Computing Machinery.

[49] R. Wash. Folk models of home computer security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, New York, NY, USA, 2010. Association for Computing Machinery.

[50] A. Whitten and J. Tygar. Why Johnny can't encrypt: A usability evaluation of pgp 5.0. In *Security Symposium*, volume 348, pages 169–184. USENIX, 1999.

[51] G. Wilson, H. Maxwell, and M. Just. Everything's cool: Extending security warnings with thermal feedback. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2232–2239. ACM, 2017.

Website	Correct Assessments	Accessibility	Perceived					
			Task A		Task B		Task C	
			Ease	Conf.	Ease	Conf.	Ease	Conf.
Gmail	5/6	4.2	3.50	4.17	4.00	4.33	3.67	4.67
Amazon	6/6	3.5	4.17	4.50	3.83	5.00	2.83	3.67
CCNIB	**0/6**	4.3	4.33	5.00	4.33	4.83	4.17	4.00

Table 2: Phase 2 results. Number of correct assessments for whether the site was legitimate or fraudulent; mean Likert scale ratings (out of 5) for the site's perceived accessibility, self-reported ease of completing the task and level of confidence in completing task in its entirety.

A Post-Task Questionnaire

Questions 1, 2, and 3 were asked after completing each task. Questions 4, 5, and 6 were asked after completing all tasks for a website.

Q1: Is this website...Legitimate or Fake?

Q2: How easy or difficult was it to complete the task? (1. Extremely difficult, 2. Difficult, 3. Neither easy nor difficult, 4. Easy, 5. Extremely easy)

Q3: How confident are you that you completed the task? (1. Extremely unsure, 2. Unsure, 3. Neither sure nor unsure, 4. Sure, 5. Extremely sure)

Q4: How would you rate the website's accessibility? (1. Extremely inaccessible, 2. Inaccessible, 3. Neither accessible nor inaccessible, 4. Accessible, 5. Extremely accessible)

Q5: How does this activity compare to your experiences with similar tasks outside of this study?

Q6: What other steps might you take if you were faced with a similar situation in real life?

B Post-Test Questionnaire

Q1: Rate your level of concern with the following digital threats on a scale of 1 (very unconcerned) to 5 (very concerned). *Ordering of options randomized per participant.*

- Someone stealing your identity
- Someone gaining access to your financial information
- Someone stealing private information about you/your family
- Your personal information being made public
- Falling victim to an online scam or fraud
- Someone hacking in to your email
- Unintentionally installing malicious software
- Your device becoming infected with a virus or malware
- Your device becoming infected with key-stroke logging software
- Someone eavesdropping on you
- Someone watching your interactions without you knowing

Q2a: Rate the effectiveness of the following protective actions on a scale of 1 (not effective at all) to 5 (extremely effective). *Ordering of options randomized per participant.*

- Frequently update software and systems
- Enable automatic updates
- Use software from official, trusted sources
- Use antivirus software
- Use strong passwords
- Use unique passwords between different sites
- Use multi-factor authentication methods
- Use a password manager
- Only use websites that include "HTTPS" in the URL address
- Think before clicking a link
- Do not open unexpected attachments

Q2b: On a scale of 1 (extremely unlikely) to 5 (extremely likely), how likely are you to take the protective actions? *Refer to listed options above. Ordering of options randomized per participant.*

C Post-Test Interview

Q1: Tell me more about what happens when you face... *an obstacle we observed or the user mentioned.*

- **Q1a:** What do you think caused this issue?
- **Q1b:** How did this problem affect your mood?
- **Q1c:** How do you think this problem affected your security or privacy?

Q2: How often do you consider your personal security and privacy when surfing the web? (1. Never, 2. Very rarely, 3. Rarely, 4. Occasionally, 5. Very frequently, 6. Always)

Q3: How safe do you usually feel when offering sensitive information online? (1. Extremely unsafe, 2. Unsafe, 3. Neither safe nor unsafe, 4. Safe, 5. Extremely safe)

Q4: If any, what are your most pressing concerns when browsing online?

Q5: What makes you feel safe online?

ID	Sex	Age	Visual acuity	Visual field	Light perception	Occupation	OS	Accessories	Assistive	Settings	Browser	Website
U01	F	20	Very limited	Very limited	Very limited	Student	W	K	JAWS	Default	IE	Gmail
U02	F	55	Somewhat	Somewhat	Somewhat	Unemployed	W	K, M, D, glasses	ZoomText	Voice reader	IE	Amazon
U03	M	26	Very limited	Somewhat	Not at all	Student	W	K, M, D	JAWS	Default	IE	Gmail
U04	M	6 3	Very limited	Very limited	Very limited	Retired	iOS	None	VoiceOver	Default	Safari	Amazon
U05	F	51	Very limited	Very limited	Somewhat	Technologist	W	K	JAWS	Default	IE	Gmail
U06	F	54	Somewhat	Very limited	Somewhat	Unemployed	W	K, M, D	ZoomText	High-contrast, voice reader	IE	CCNIB
U07	M	51	Very much	Very limited	Somewhat	Contractor	W	K	JAWS	Default	IE	Amazon
U08	M	41	Somewhat	Somewhat	Somewhat	Unemployed	W	K, M, D, magnifying glass	ZoomText	Default	IE	Gmail, CCNIB
U09	F	68	Somewhat	Not at all	Somewhat	Small business owner	iOS	Book stand	None	Default	Safari	Amazon
U10	M	68	Very limited	Very limited	Not at all	Retired	W	K	JAWS	Default	IE	CCNIB
U11	F	70	Somewhat	Somewhat	Somewhat	Retired	W	K, M, D	ZoomText	Default	IE	Gmail, CCNIB
U12	M	51	Very limited	Very limited	Somewhat	Unemployed	W	K, M, D, glasses	None	High-contrast, cursor enlarge	Chrome	Amazon, CCNIB
U13	M	40	Somewhat	Not at all	Somewhat	Unemployed	W	K, M, D	ZoomText	Default	IE	Gmail, CCNIB
U14	M	55	Very limited	Very limited	Very limited	Customer Service	W	K	JAWS	Default	IE	Amazon

Table 3: Participant demographics and the devices/software used during user study sessions. OS column represents the operating system used where: “W” represents Windows 10. Accessories column represents the technology used during the session where: “K” is keyboard with tactile markers, “M” is standard computer mouse, and “D” is display monitor.

Code	Description	Code Group
Personal abilities or attributes	Participant expresses confidence or apprehension in their abilities/attributes which play a role in completing tasks	**Personal abilities and attributes**
Preferences	Participant expresses preference (or aversion) for certain techniques to completing tasks.	
Usability or accessibility obstacles	Instances mentioned during discussion or experienced while completing tasks in which participants face challenges that infringe usability/accessibility	**Usability, accessibility obstacles**
Guesswork	Participant hypotheses in how the system works, what it is doing, or how to strategize interactions to achieve desired ends.	Guesswork
Obstacle compensations or workarounds	Participant techniques in overcoming issues while trying to complete tasks.	Obstacle compensations or workarounds
Mental models of websites	Participant understandings of how websites work based on experiences and expectations.	**Mental models of websites**
Learned functionality language	Terms and phrases indicating participants' unique interaction with and navigation of websites, using assistive technology or software.	Learned functionality
Roles of website expectations	Participant expectations of a site/system and their implications on task processes.	Role of expectations
Legitimate and/or secure websites	Cues which participant uses to validate website legitimacy or security.	
Apathy towards privacy/security	Instances in which participant is unconcerned for their online privacy/security.	**Security and privacy attitudes**
Security is secondary	Participant gives higher priority to other aspects of the interaction than their personal privacy/security.	
Security uncertainty	Participant expresses uncertainty in maintaining their personal privacy/security.	
Security concerns	Participant describes their privacy/security concern(s).	
External influences	Evidence of brand, institution, software, etc. influence on participants' understanding of privacy/security.	External influences
Tolls of infringed security	Participant describes negative consequences (experienced or presumed) of privacy/security infringements.	
Incomplete/inaccurate security mental models	Participant techniques in protecting themselves which are based in incomplete/inaccurate understandings of threats.	Understanding of threats
Security management	Participant describes their methods in managing their personal privacy/security.	**Security management techniques**

Table 4: Final version of the codebook which describes participants' interview data and other feedback.

WebAlly: Making Visual Task-based CAPTCHAs Transferable for People with Visual Impairments

Zhuohao Zhang[1], Zhilin Zhang[1], Haolin Yuan[2], Natã Barbosa[1], Sauvik Das[3], Yang Wang[1]
[1]University of Illinois at Urbana-Champaign [2]John Hopkins University [3]Georgia Tech
{zhuohao4, zhilinz2, natamb2, yvw}@illinois.edu, {hyuan4}@jhu.edu, {sauvik}@gatech.edu

Abstract

Task-based visual CAPTCHAs are a significant accessibility hurdle for people with visual impairments (PVIs). What if PVIs could transfer task-based visual CAPTCHAs to a helper to solve? How might PVIs want such a system configured in terms of from whom they would solicit help and how they would compensate this help? To answer these questions, we implemented and evaluated a proof-of-concept assistive transfer system — WEBALLY — that makes task-based CAPTCHAs transferable by allowing PVIs to source just-in-time, remote control help from a trusted contact. In an exploratory, role-play study with 10 pairs of participants — a PVI and a friend or a family member — we asked participants to use WEBALLY in four different configurations that varied in source of help (friend vs. stranger) and compensation (paid vs. volunteer). We found that PVIs liked having WEBALLY as an additional option for solving visual CAPTCHAs, when other options that preserve their independence fail. In addition, many PVIs and their friends felt that using the system would bring their relationship closer. We discuss design implications for transferable CAPTCHAs and assistive transfer systems more broadly, e.g., the importance of complementing rather than replacing PVIs' existing workflows.

1 Introduction

Motivation. Large swathes of the web remain inaccessible for the 285 million people with visual impairments (PVIs) [41]. For instance, CAPTCHAs (Completely Automated Public Turing tests to tell Computers and Humans Apart) are commonly used to authenticate users in numerous day-to-day web surfing tasks [33] (e.g., registering new accounts, leaving comments on social media, and completing financial transactions), yet were rated as the most problematic item for PVIs in a global study by WebAIM [51].

CAPTCHAs are inaccessible for PVIs because they require users to engage in complex visual-processing tasks. For example, today, task-based visual CAPTCHAs, such as Google reCAPTCHA [46] and GeeTest [21], are widely used across the web and require users to perform high-precision operations such as selecting a subset of images from the gallery, or dragging a slider to solve a puzzle. These CAPTCHAs are challenging if not impossible for PVIs to solve independently.

Existing solutions. Prior work has explored a number of solutions to make CAPTCHA-solving easier and more accessible. Most commonly, PVIs use audio CAPTCHAs instead. However, prior work has found that audio CAPTCHAs are disproportionately hard for PVIs relative to how hard visual CAPTCHAs are for people without visual impairments — they are significantly slower and require more attention and memory-capacity [15]. Other solutions to help PVIs in the short-term include automated CAPTCHA solving services (e.g., WebVisum [56]), but these solutions work only for simple visual CAPTCHAs in which people are asked to identify distorted letters and numbers. There are also CAPTCHA solvers (e.g., Anti-CAPTCHA [7]), but these pose security risks — they require users to install software with dangerous system-level permissions. In short, while existing solutions do help PVIs solve or bypass CAPTCHAs in some cases, there are still many other cases in which they fail PVIs.

To help PVIs in overcoming day-to-day web accessibility hurdles outside of CAPTCHAs, crowdsourcing and friendsourcing methods have shown great promise. Many PVIs use remote assistance services or ask friends around them to directly help [58]. While these methods require interdependence, crowdsourcing and friendsourcing can help PVIs with these accessibility challenges that they might encounter in today's web. Prior art, such as BeMyEyes [10] and VizWiz [16], have explored connecting PVIs in need of help with remote assistance from sighted helpers to, for example, answer ques-

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

tions about surroundings or request for vocal-guidance on using inaccessible interfaces. However, these applications are limited to descriptive guidance, which limit their utility in solving task-based CAPTCHAs.

There are also trade-offs to sourcing help from the crowd or from friends. While crowdsourcing might pose privacy and security risks for remote control — in which a crowd helper (stranger) takes over the PVI's system to solve CAPTCHAs on their behalf — friendsourcing better aligns with existing workflows for PVIs: PVIs often seek assistance from their friends and family members to overcome accessibility challenges in the physical world [1]. However, friendsourcing can also lead to privacy problems as users might not want to expose their ongoing activities to friends. Friendsourcing can also reduce monetary cost of using paid crowd workers. However, friendsourcing can be slower and less reliable than crowdsourcing [9], and PVIs might want to avoid burdening their social connections with frequent requests for help [47].

Research questions. In short, PVIs commonly encounter task-based visual CAPTCHAs that frustrate and encumber their use of the web, yet existing solutions fall short of supporting their needs. With an overarching goal of designing inclusive privacy/security tools [55], we posit that an *assistive transfer system* that allows PVIs to solicit just-in-time help where a willing helper directly solves an outstanding reCAPTCHA challenge on the PVIs' behalf may be beneficial, if seen as a "last resort" when other options to retain independence have been exhausted. However, the design space of such assistive transfer systems have not yet been explored, and there are a number of open questions regarding how PVIs might use such a tool and how it might be configured — e.g., whether and in what contexts the help should be sourced from friends or strangers and whether they would prefer the help to be free or paid. To bridge this gap, we designed and implemented WEBALLY, a proof-of-concept assistive transfer system that allows PVIs to transfer the solving of task-based CAPTCHAs to others, and used WEBALLY as a design probe in an exploratory user study to answer these research questions:

- **RQ1:** What are both PVIs' and helpers' general impressions towards using assistive transfer systems like WEBALLY?
- **RQ2:** What are the perceived privacy and security risks for PVIs when transferring CAPTCHAs to others?
- **RQ3:** What factors influence PVIs' preferences in configuring from whom to source help?
- **RQ4:** What factors influence PVIs' preferences towards compensating helpers?
- **RQ5:** What is the perceived impact, of using assistive transfer systems like WebAlly, on the social relationship between PVIs and friends from whom they solicit help?

User study. We conducted a within-subjects, two-by-two (crowdsourcing vs. friendsourcing, paid vs. free) lab study (over Zoom video conference) with 18 participants (10 PVIs and 8 sighted friends) to answer our research questions. We recruited participants in pairs (one PVI and one friend who served as a remote helper), and had PVIs use WebAlly to request their helper to solve Google reCAPTCHAs for them. To simulate the different configurations of our study, we had participants role-play — a technique commonly employed in usable security research [48]. For example, while each PVI participant had the same helper (their friend) in all conditions, we had participants envision themselves in a situation where they would need to solicit help from the crowd/stranger or a friend, and where they would need to pay for this help or not. Our data included notes taken by researchers as participants engaged in the study tasks, and post-study exit interviews. Given the limitations of our study, we note that our key contribution is *less* about WEBALLY and its evaluation in and of itself, but *more* about — the knowledge of designing transferable CAPTCHAs and assistive transfer systems more broadly — gained through the design and evaluation of WEBALLY with stakeholders, as is common in HCI design research [63].

Findings. We found that while transferring CAPTCHAs requires interdependence, both PVIs and helpers appreciated having a system as a last-resort alternative to other accessibility solutions (e.g., audio CAPTCHAs). PVIs felt that WEBALLY could mitigate the privacy and security risks entailed by transferring a task to helpers by intelligently cropping to only task-relevant parts of the PVIs' screen, but some still had concerns about sourcing help from others altogether. Helpers, on the other hand, had some concerns over requests from strangers in the form of an open link. We also uncovered four important factors that may affect PVIs' perceptions towards using an assistive transfer system for CAPTCHAs: the type of webpage that embeds CAPTCHA (e.g., whether it contains private browsing data), the use case (e.g., whether it is financial-related), helper availability, and impact on requester-helper relationships. In addition, we also explored PVI users' personal preferences towards compensating the helpers.

Contributions. Our work has two main contributions: (1) We introduced and explored the design space of assistive transfer systems for task-based CAPTCHAs, and (2) We implemented a proof-of-concept assistive transfer system, WEBALLY, and conducted an exploratory evaluation with both PVIs and helpers to synthesize design insights for assistive transfer systems in the context of Google reCAPTCHAs.

2 Related Work

2.1 PVI with CAPTCHAs

CAPTCHAs are designed to distinguish humans from robots. They deter hackers from abusing online services and are served millions of times a day [30]. Traditional CAPTCHAs

usually pose a visual challenge, like recognizing images, words, or numbers out of specific images. These interactive tasks are meant to be simple for human users. However, CAPTCHAs are notoriously inaccessible for folks with visual, physical, cognitive, or auditory disabilities [14,26,37,49]. For PVIs who use screen readers, specifically, visual CAPTCHAs pose a hurdle that often cannot be overcome without relying on other external help — be it a friend or a service designed to bypass CAPTCHAs.

As an alternative, audio CAPTCHA is more accessible for PVIs. However, audio CAPTCHAs are not always available on many websites, and current audio CAPTCHA designs have proven to be difficult and time-consuming for PVIs throughout several research studies [15, 32, 34, 35]. To improve security against speech recognition algorithms [17], the audio file provided to users are usually speakers saying words at randomly spaced intervals with background noise. These interferences challenges both automated agents and human users [32, 52]. Many existing research studies have also tried to increase the accessibility of audio CAPTCHAs. Fanelle et al. designed four novel audio CAPTCHAs to increase accuracy and speed [24]. Jain et al. proposed *reCAPGen*, a system that uses automatic speech recognition for generating more usable and secure audio CAPTCHAs [32]. They all explored how users (especially PVIs) can independently solve audio CAPTCHAs. Prior work also provided many examples of directly breaking CAPTCHAs. Some early research leveraged image and pattern recognition techniques to break visual CAPTCHAs [20, 38, 60]. More recent research also provided various types of hackings towards task-based CAPTCHAs like Google reCAPTCHA [8, 36, 50, 61]. There are also many paid CAPTCHA solving services like Anti-CAPTCHA [7] and Buster [25]. Although these techniques could be easily adopted in browser extensions or system-level applications to hack CAPTCHAs directly for PVIs, the original purposes of these research are still aimed for improving the CAPTCHA's security by revealing how they can be hacked. Additionally, using hacking services like Anti-CAPTCHA would introduce privacy and security issues. Users will need to download browser extension files directly from their website rather than installing from official stores, and users are required to edit their computer's registry to make the tool work.

2.2 Privacy and Security Concerns of PVIs

As online resources have been more and more available and accessible for users with visual impairments, there is a trend towards empowering PVIs to protect private information and their online security. Gurari et al. introduced the first visual privacy dataset originated from PVIs, revealing a challenge of understanding and protecting their privacy needs [27]. The dataset also includes information that can be easily captured on PVIs' computer screens. As crowdsourcing remote assistance services like BeMyEyes and Eyecoming [42] have been widely used by PVIs and make their lives easier, researchers have also investigated how these services would raise privacy and security risks [3, 5, 6, 59]. Akter et al. conducted a study to understand privacy concerns when PVIs use camera-based assistive technologies [6]. Ahmed et al. took another angle and studied the information sharing preferences of sighted bystanders of assistive devices [3]. Existing research has also shown that PVIs have strong security and privacy concerns in using CAPTCHA [2, 4, 22, 28, 31]. Holman et al. identified their top 10 security challenges and CAPTCHA has been listed as the top one challenge [31], which poses a challenge of how to help PVIs solve these small tasks like CAPTCHA without compromising their privacy and security.

2.3 Sourcing help for PVIs

Socio-technical researchers conducted many studies on collaborative systems. Traditional crowdsourcing has proved a convenient way to get answers quickly from the crowd. The VizWiz smartphone application allows visually-impaired users to send visual questions to sighted crowd workers and get answers soon [16]. However, such services can be limited due to the cost of the paid crowd workers, which might add extra and unexpected burden to PVIs [19]. Friendsourcing could also help users solicit answers and assists from friends via online social network services, and the answers are often from more trustworthy and tailored to their interests than using a search engine [40]. Traditional online social network sites used for these include Facebook and Twitter [13, 39, 40, 47]. For PVIs specifically, AbdraboTarek et al. proposed an assistive tool for blind users to friendsource help for daily activities via smartphone and Twitter [1]. Brady et al. studied PVIs' perceptions of social microvolunteering via Facebook answering visual questions on behalf of blind users [19].

Crowdsourcing and friendsourcing have their unique advantages and disadvantages in many aspects. For example, differences exist about compensation and response rates and potential impact on social relationships. Zhu et al. studied the effects of extrinsic rewards and monetary payments to further investigate how friendsourcing would impact PVIs' social relationship with their friends [62]. Other research also revealed how these rewards might undermine the original motivation that drives friendsourcing activity and change the perceived relationship between people [29, 43, 53, 54]. In addition, independence is often considered as a goal in assistive technologies [11]. Even sometimes the goal is not explicitly stated, the researchers agree that "all accessible computing approaches share a common goal of improving independence, access, and quality of life for people with disabilities" [57]. However, as Bennett et al. pointed out, interdependence is also valuable because the interactions between people with disabilities and their allies are often two-way and mutually beneficial [11]. In our work, we aimed to use a novel collaborative method as a probe to explore these different design spaces. We also looked at both PVI user side and helper's side,

which has not yet been fully investigated by other researchers.

3 Design Considerations

We began by identifying the challenges PVIs experience with task-based CAPTCHAs — we enumerated existing solutions to help PVIs overcome these CAPTCHAs, investigated how existing solutions fall short of PVIs' needs, and uncovered how this transferable method could play a role in helping PVIs solve these tasks. Then, we synthesized several design goals to support PVIs with an accessible tool to solve Google reCAPTCHA, one of the most commonly used task-based CAPTCHA on the web.

3.1 Design Challenges

C1 - Providing PVIs with more direct manipulation Existing tools such as BeMyEyes and Eyecoming provide PVIs with remote assistance services: e.g., providing descriptive guidance on how to operate an interface and navigation guidance via smart glasses. However, these collaborative assistance services are limited to providing indirect help; PVIs rely on helpers' textual or verbal guidance, either synchronously or asynchronously, to solve the task on their own. While this type of assistance is helpful, fosters independence, and is widely used by PVIs, it can become challenging when the task requires precise hand-eye coordination (e.g., moving the mouse and clicking specific areas). One opportunity to address this challenge is to afford a remote helper direct control of the PVIs system to solve the task-based CAPTCHA on behalf of the PVI. However, it is still challenging to make remote control assistance secure and accessible for PVIs.

C2 - Protecting privacy and security while helping While remote control assistance could help PVIs overcome task-based CAPTCHAs, it might also bring privacy and security issues: as many PVIs may be unable to receive visual feedback or otherwise monitor helpers' behaviors, remote helpers could perform malicious actions on PVIs' devices without their awareness. The helper could also become aware of what the PVI is trying to do online, or be able to see sensitive personal information that may be present on the PVIs screen. Thus, there is a need to protect PVIs' privacy and security as they receive help through remote assistance systems and services — both for indirect descriptive guidance (when PVIs usually need to point their cameras to the computer screen) and direct remote control.

C3 - CAPTCHA restrictions In CAPTCHA design, there is usually a trade-off between security and user experience. In achieving its original purpose of differentiating humans from bots, task-based CAPTCHA can be challenging even for humans as a result of making it more robust against bots. For example, Google reCAPTCHA, the most common type of task-based CAPTCHA, has a solving time limit of two minutes. It also expires within one minute before submission, such that users must complete and submit a form protected by reCAPTCHA before it expires, lest they have to solve another reCAPTCHA challenge. Remote assistance services usually take more than two minutes to post requests, find volunteers, synchronize with the helper, get help, and get notifications when the session is complete. Often the case is that PVIs need to wait for someone to answer their requests. It is naturally challenging to source help in a short amount of time before the current CAPTCHA expires.

3.2 Design Goals

To address these challenges, we highlighted several goals that we identified as essential for designing an efficient and accessible CAPTCHA-solving tool for PVIs. Our high-level design goals were to integrate social support, reduce human effort, source help efficiently, protect PVIs' privacy, and still maintain the security utility of CAPTCHAs — differentiating between humans and bots.

G1 - Limited remote control To provide PVIs with more direct help and maintain their privacy and security at the same time, our goal is to design a limited, sandboxed remote control system in which helpers are restricted in the actions they can perform and the screen information they can see. Specifically, helpers should only be able to perform actions necessary to solve the CAPTCHA, and should only be able to see parts of the PVI's screen that is relevant to the CAPTCHA. In this case, helpers cannot access any sensitive information or perform other actions on the PVIs' personal devices.

G2 - Simplicity of use Task-based CAPTCHAs like Google reCAPTCHA have time limits, after which they expire and a new visual challenge is issued. One typical assistive system often includes: (1) send the PVIs' request, (2) synchronize with a helper, (3) have the helper complete the task. For a fast and simple solving experience, the user interface for the helpers should be simple and straightforward so they can minimize the completion time to avoid expiration.

G3 - Accessible in usage To ensure that our remote assistance tool is accessible, our goal is to make all its functions available via keyboard shortcuts. Another design goal to enhance accessibility is to provide audio feedback at every stage of helping, to notify PVIs about the current state of the task and what the helper is doing.

3.3 System Design

Guided by design goals, we implemented WEBALLY — a proof-of-concept system that connects PVIs with their friends when they encounter Google reCAPTCHAs. WEBALLY creates an interactive screenshot of the reCAPTCHA and sends it to the helper. This interactive screenshot serves as a canvas on which helpers' can perform actions to solve the reCAPTCHA (e.g., clicking on tiles, dragging UI elements); these actions, in turn, are reflected on the PVIs screen. However, access to the

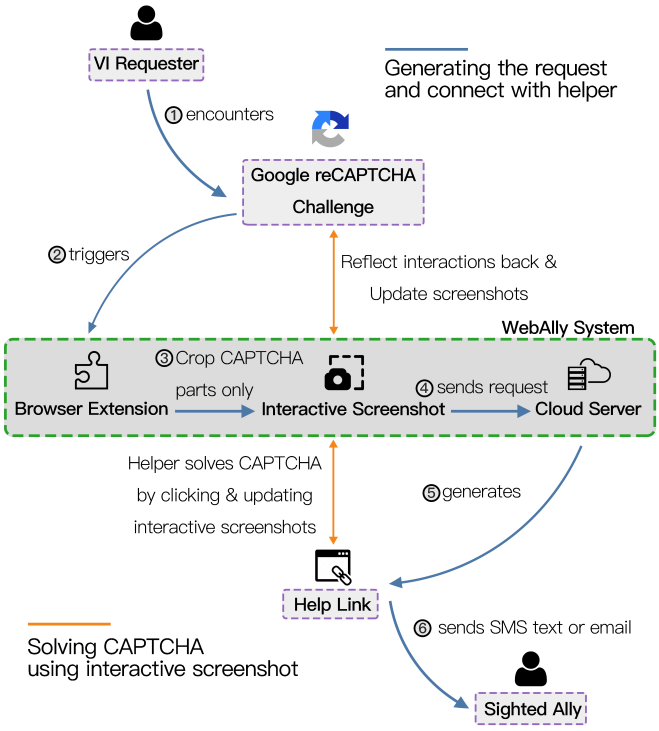


Figure 1: Workflow of WebAlly: PVI triggers the tool and sends the request, and the tool builds a channel between the PVI and the helper via an interactive screenshot: Reflect helper's clicking on requester's screen and update screenshots back to helper's interface until finished.

source device is totally reconstructed — helpers see only task-relevant screen information, and only pre-specified interactions (e.g., clicks) will be reflected and simulated on the PVI's screen. In the following sections, we will use the terminology "requester" to refer to users with visual impairments sending requests for help, and "helper" to refer to their friends and family offering help. The WEBALLY system stores cropped screenshots and helpers' contact information only temporarily — i.e., for the duration of the transfer task.

3.3.1 System Overview

WEBALLY is implemented as a browser extension, written in JavaScript and executable on Chromium-based browsers (such as Google Chrome, Microsoft Edge, Opera and Brave). We also incorporated OpenCV to pre-process images, and WebSocket as a channel to transmit messages in real time. The workflow contains a one-way request from the requester and a synchronous collaboration process between the requester and the helper (see Figure 1). We have open-sourced the source code for WEBALLY [1].

[1] https://gitlab.engr.illinois.edu/salt-lab/webally

3.3.2 Interface Details

The requester interface is simplified to be accessible for PVIs. First, the requester will enter and store the contact information for a helper using a keyboard shortcut to activate the function. The requester can then activate the extension and send the request to the preset helper using an editable keyboard shortcut. The WEBALLY system then takes a screenshot of the current browser tab and uses the template-matching feature in OpenCV.js to crop the screenshot down to just the region that contains the Google reCAPTCHA task. Only the cropped image will be sent to the helper. If the reCAPTCHA task is successfully solved, the requester can end the collaborative session and continue the task at-hand, e.g., submitting a form. If the helper failed to pass the test in time, the requester could ask again via the same keyboard shortcut.

The helper will receive SMS texts or emails with a URL to a secure page. The helper can then complete the requested task on their own browser. The helper's interface contains instructions on what to do and the cropped interactive screenshot. To reflect the helper's actions on their screen into the requester's screen, WEBALLY record, transmit, and simulate the helper's clicks on the interactive screenshot via WebSocket. As the reCAPTCHA interface updates the tile images after each click, the system also detects the updated part, crop it using OpenCV, and reflect these changes back on the interactive screenshot presented to helpers. Thus, the helper's experience mimics how they might solve a CAPTCHA for themselves.

4 User Study

While WEBALLY is fully functional and can be used in practice, it is not a finalized product — it is a design probe that we presented to PVIs in order to model their perceptions of and configuration preferences for transferring reCAPTCHA challenges to remote helpers. To that end, we conducted a two-by-two within-subjects study with 18 participants (10 PVIs and 8 helpers) to evaluate WEBALLY and answer our research questions. Our study was IRB-approved. We had PVIs and their helpers roleplay using WEBALLY to transfer a Google reCAPTCHA task in four different scenarios that varied in source-of-help (stranger or friend) and compensation strategy (paid or voluntary). While we report on some descriptive quantitative findings, we note that our main findings are qualitative — our goal was less to comparatively evaluate different conditions, and more to understand PVIs perceptions of WEBALLY under different configurations.

4.1 Participants

We recruited participants in pairs: one PVI and one helper who the PVI considered a close friend or family member. We recruited 18 participants (see Table 1), including 10 participants (6 males, 4 females) with visual impairments (referred to as requesters R1-R10) and 8 participants (3 males, 5 females)

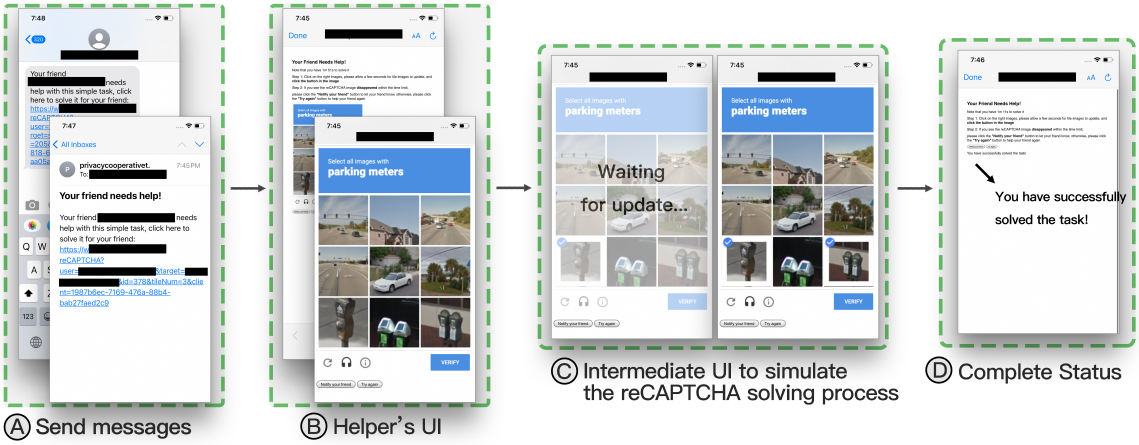


Figure 2: User Interface from the helper's side: Requester sends message, helper opens the link, solve the CAPTCHA as usual, and click "VERIFY" to complete (The solving process might take multiple visual challenges to pass).

Table 1: Participant demographics, including requesters' age range, gender identity, self-described visual ability, screen reader of use, and their relationship with the helpers.

PID	Age	Gender	Visual ability	Screenreader	Helper	Gender	Relation
R01	25-34	Male	Blind	NVDA	H01	Male	Brother
R02	18-24	Male	Blind	NVDA	H02	Female	Sister
R03	55-64	Female	Deafblind	NVDA	H03	Male	Brother
R04	25-34	Female	Blind	JAWS	H04	Female	Brother
R05	25-34	Male	Blind	JAWS	H05	Female	Friend
R06	25-34	Male	Low Vision	NVDA	H06	Female	Friend
R07	25-34	Male	Low Vision	NVDA, JAWS	H07	Male	Friend
R08	35-44	Male	Blind	JAWS	H08	Female	Friend
R09	45-54	Female	Blind	JAWS	H08	Female	Friend
R10	35-44	Female	Blind	JAWS	H08	Female	Friend

without visual impairments (referred to as helpers H1-H8 just for simplicity). Among PVIs, one self-identified as deafblind, two as low-vision, and seven as blind. One helper (H3) self-identified as having some hearing impairment. For the screen readers they were using, one was using both NVDA [45] and JAWS [44], four were using NVDA, and five were using JAWS. For the study sessions, requesters R1-R7 were in pairs with helpers H1-H7 accordingly in our study, and H8 was a mutual friend to requester R8-R10, who joined the study with them for three times. For their relationship, one helper is the sister, three helpers are the brothers, and four helpers were friends of the corresponding PVIs.

4.2 Apparatus

We used the WEBALLY prototype to conduct the study. We asked requesters to install WEBALLY on their Chrome browsers before the study. Instead of using a real website for testing WEBALLY, we used a demo website (https://www.google.com/recaptcha/api2/demo) which contains a login form simulating what users would encounter in real settings. The reason we are not using real websites is that,

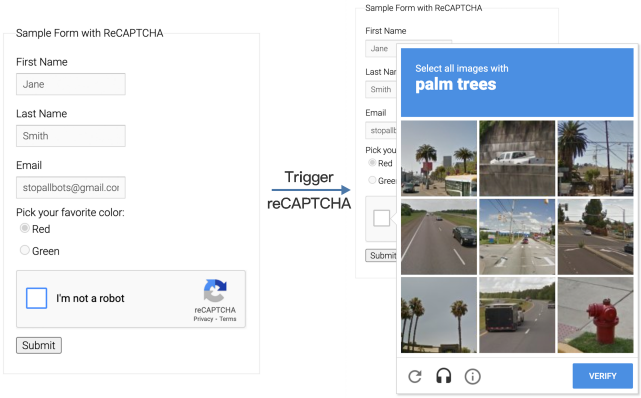


Figure 3: The demo website to trigger Google reCAPTCHA.

unlike real websites, the demo website is designed to *stably trigger* the Google reCAPTCHA task to ensure PVI users would face this particular task in the study.

For helpers, we asked them to have a device (either computer or mobile devices) through which they could join the meeting and receive SMS messages and/or emails.

4.3 Procedure and Data Collection

We conducted a lab study to explore the design space of WEBALLY. While a field study could be helpful in evaluating near-final design concepts in ecologically valid contexts, a lab study allowed us to capture rich, qualitative data on users' acceptance of and perceived feelings towards WEBALLY under different experimental configurations in an exploratory and controlled way. We do not see WEBALLY as a near-final design concept. Rather, it is a design probe and sensitizing

concept that we implemented to evaluate with stakeholders in order to synthesize design knowledge, as is common in HCI design research [63]. As such, we elected to run a lab study.

The participant sessions were all conducted remotely via video conference calls. Before the sessions, we confirmed that participants had regular access to a laptop or desktop computers, and that they used Mozilla Firefox or Google Chrome to surf the web in their daily life. All sessions lasted about an hour, including a post-study interview to learn the requesters' and the helpers' feedback separately.

At the beginning of each study session, two researchers and two participants (the requester and the helper) would join the online meeting. Participants were introduced to the purpose of the study and the study procedures. In some cases, the requesters did not or failed to install the tool in advance. The researchers then helped them install the tool via the online meeting and guided the requesters to type the helper's information into the WEBALLY's interface.

Scene Setup and Role Explanation

After the preparation and installation, the researchers would divide the two participants into two breakout rooms to simulate remote collaboration (i.e., they did not need to be physically co-located to use the system). The two researchers who helped conduct the study went into the each of the two breakout rooms in order to observe helper and requester behaviors and answer their questions. Splitting requesters and helpers up into two separate breakout rooms helped approximate a real-world scenario in which a PVI would need to to solve a Google reCAPTCHA task and choose to source remote help from friends. After the researchers and the participants settled in different rooms, the researcher in the requester room (referred to as Researcher 1) introduced the tasks and asked the participants questions from a pre-study questionnaire (A.1) about their experience with CAPTCHAs and Google reCAPTCHA. The questionnaire was designed to understand how participants generally solve reCAPTCHAs and the challenges faced in solving these task-based CAPTCHAs.

After asking participants about their prior experiences with CAPTCHAs, researchers explained the scene and role setup. In the study, we asked helpers to play the role of both friend and stranger (i.e., crowd worker). This ruse was made more believable by the fact that requesters and helpers were separated from one another during the study in order to simulate the remote collaborative setting. As such, our participants did not necessarily know the exact identity of who might have been helping them or requesting their help for a given reCAPTCHA task. PVI requesters were told that they will transfer the request to either their friend or an unknown crowd worker in different scenarios. Helpers were told that they will receive a request from either their PVI friend or a stranger who is also using our tool. Thus, while we employed role play, participants had reason to believe their roles were, in fact, true — strengthening ecological validity, though still a limitation.

Similarly, to strengthen ecological validity from the helpers' perspective, helpers were instructed that the requests from their PVI friends or a stranger may come at any time, and that they could do whatever they pleased in the meanwhile rather than waiting for WEBALLY requests. When their assistance was requested, they would be notified via SMS or email — just as they would in real settings when they are not necessarily prepared to help their friends exclusively.

After explaining scene and role setups, the lab study began. Researcher 1 asked the requester to imagine that they are under one of four different scenarios. We had a 2 x 2 within-subjects experimental design with two factors: Source of help (ally vs. stranger), and compensation for help (free vs. paid). In the helper's room, the researcher (referred to as Researcher 2) also introduced the different configurations to the helper and asked them to behave accordingly under different scenarios (e.g., imagine that a blind person whom you do not know asks your for free, voluntary help to solve a reCAPTCHA).

In the study, the researchers randomized the order in which the four configurations were presented to reduce order effects. Under each configuration, researchers asked participants to use WEBALLY to solve a reCAPTCHA task together. Broadly, the PVI asks their helper for assistance, the helper solves the reCAPTCHA on his/her own screen, which would be automatically transmitted back to the helper through the limited remote control functionality we implemented in WEBALLY. To reliably trigger a reCAPTCHA challenge under each configuration, requesters used the aforementioned demo website. Then the requester would send the challenge to the helper through a keyboard shortcut. The helper could receive the help request message via SMS or email (see Figure 2 A). The message would contain a broad description of the requested remote assistance task (see Figure 2 D) and a web link containing the interactive screenshot of the Google reCAPTCHA (see Figure 2 B, C). In ideal scenarios, each configuration contained just one task-solving process. However, if the helper failed to complete the task within the time limit and the reCAPTCHA expired, the researchers asked participants to repeat the process again under the same configuration.

After completing all of the four configurations, the researchers conducted an exit interview asking about participants' detailed experiences with the system and their preferences among the four configurations. The questions include general feedback, suggestions on improvement, which configuration they chose as their favorite and why, their privacy and security concerns in details, and how the request might alter the relationship dynamics between requesters and helpers.

4.4 Analysis

Upon participant consent, we video-recorded all the online sessions and took notes from the procedure and the interview.

We transcribed the videos and used thematic analysis [18] to qualitatively analyze the study. Broadly, our analysis was driven by our core research questions and covered perspectives of both PVIs and helpers. For PVIs, our codes cover prior experience with solving CAPTCHAs (e.g., tools used or methods for sourcing help), their general feelings about WEBALLY, concerns on using WEBALLY and its potential privacy/security risks, their nuanced context-based preferences with respect to crowd/friendsourcing and paid/unpaid versions of WEBALLY, their desire to use WEBALLY in the future (willing to install and recommend), and their perception of how WEBALLY might affect their relationship with friends and family. For helpers, our analysis covered their overall feeling and suggestions, willingness and general availability to help PVIs or even a broader user group with/without compensation, their choice on different configurations to help, and related feelings.

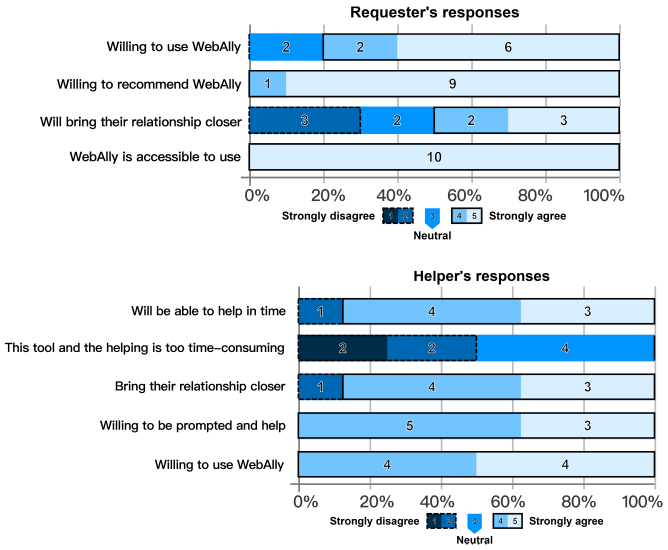


Figure 4: Requesters' and helpers' responses on the acceptance of the tool, possible change in their relationship, and the tool's accessibility level.

5 Results

5.1 RQ1: General Impressions and System Performance

After participants used WEBALLY under all study conditions, we asked about their general impressions of using the system.

All participants reported that they liked the tool but also identified its pros and cons. For example, many (9 out of 10 requesters and 7 out of 8 helpers) thought the solving process was effective, and they were excited about the overall idea of transferring complex CAPTCHAs to others and how the tool was easy and effective. For instance, R1 praised WEBALLY as a *"really bright idea."* R2 said *"really appreciate[d] this invention"* because *"many websites do not have audio CAPTCHAs, and many audio ones just do not work."* Many helper participants also thought positively about the tool. For instance, H6 commended that *"it is simple and helpful, and I feel great to help someone."* H4 cited privacy benefits: *"it is great especially from a privacy perspective"* in reference to the privacy-preserving image cropping features.

However, some users (3 out of 10 requesters and 4 out of 8 helpers) also pointed out limitations to task transferrance. Most importantly, two PVIs were concerned that their friends and family would not be available to solve a CAPTCHA before it expires in two minutes. For instance, they hesitated in sending request messages to their friends when the timing would be inconvenient — e.g., late at night. This result, while unsurprising, is a fundamental limitation to making tasks transferrable to friends and other social connections; transferrance, thus, must be considered a complement to existing accessibility solutions — a "last-resort" option that can be relied upon when all other options to retain independence have been exhausted. R8 also noted that *"it is limited to Google CAPTCHA, and it takes some time. It also expires sometimes."* Indeed, the overhead implicit in task transferrance may often be a hurdle in transferring tasks with short fuses — especially those that are dynamic and require round-trips (e.g., reCAPTCHA tasks in which tiles get updated). However, while WEBALLY was designed for Google reCAPTCHAs, it is easy to generalize the core concept of task transferrance to other accessibility challenges as well.

Among sighted helpers, some (3 out of 8) also expressed concerns about overhead and delay. H1 mentioned that *"the server response time could be improved for a bit"* and H2 said *"sometimes I need to wait for a while for images to update and I got confused when I clicked other squares too quick."* While network latency is unavoidable, WEBALLY could be improved to more gracefully handle this latency.

Participants also provided insights about WEBALLY relative to existing tools. Prior to this study, most of our PVI participants used remote assistance services or CAPTCHA-solving tools. They mentioned the similarities and differences between WebAlly and tools like BeMyEyes and WebVisum. The results showed the uniqueness of WEBALLY, which could adopt helpers' direct visual perception and corresponding action into solving task-based CAPTCHAs. Services like BeMyEyes provided a crowdsourcing solution to gather descriptive guidance. WebVisum is one of the earliest methods designed for solving visual CAPTCHAs, but many participants mentioned that it stopped maintenance long ago, and they have been relying on friends or family members physically around them to help them solve Google reCAPTCHAs.

Similar to prior work, most requesters (9 out of 10) mentioned that audio CAPTCHAs are not always accessible, and

they will still need help from a person. Based on our pre-study interview, a typical flow for PVIs to solve a Google reCAPTCHA was: (1) try to continue with their task without solving the CAPTCHA (e.g., if they were already logged into their Google accounts); (2) try to solve the audio version; and (3) if failed, ask nearby friends to solve the visual CAPTCHA for them. Half of the requesters mentioned that they could not solve the challenge if there were no friends around. This result suggests the potential value of assistive transfer systems.

We also recorded the time duration and success rate of solving CAPTCHAs as complementary data. However, it is important to note that since the helpers were present and ready to solve transferred CAPTCHAs in our lab study, the time to solve CAPTCHAs through WebAlly might be longer in practice. In our lab study, the WEBALLY tool performed well under the four different configurations. We consider a CAPTCHA task to be successful if a PVI user could pass a Google reCAPTCHA by having a helper solve the CAPTCHA via the interactive screenshot within the time limit. Overall, for all the first-time trials in all 10 study sessions, the success rate was 60%, with most failures being due to the 2-minute time restriction imposed by Google reCAPTCHA. On the first trial, helpers were not yet familiar with the tool, resulting in delays. However, as helpers gained familiarity in later trials, their speed and success improved. Indeed, the overall success rate across all task sessions was 88% (SD=0.10). Moreover, all helpers were able to solve a transferred CAPTCHA challenge the second time around if they failed the first time (likely due to their being immediately available and prepared).

Most times, the helpers could successfully solve the task within one minute. The average solving time for each task (if the task was successfully solved) was 37.9 seconds (SD=12.65). For the three study sessions (8th-10th) with the same helper (requesters R8-R10 and helper H8), the solving time decreased (27.25 and 31.5 seconds) and the success rate increased (100% for both), suggesting that trained helpers can complete the transferred tasks quickly and reliably.

5.2 RQ2: Privacy and Security Perceptions

To understand users' privacy and security perceptions of WEBALLY, we asked our participants open-ended questions about concerns that arose in their use of WEBALLY under different configurations, e.g., sourcing help from friends versus strangers.

Privacy and Security Perceptions of The Tool At first, requesters were not aware of the image cropping feature of WEBALLY because the cropped interactive screenshots were only available to the helpers. Without this knowledge, 3 out of 10 requesters (R3, R4, R8) proactively asked researchers if the helpers (irrespective of whether they were friends or strangers) could see their entire screen as would be the case with traditional remote control assistance. After learning that WEBALLY crops what helpers can see to only task-relevant portions of the UI (e.g., the reCAPTCHA challenge), all requesters found this feature useful in protecting their privacy and security.

Privacy and Security Perceptions of Transferring Tasks Participants also varied in their perceptions of transferring online tasks more generally. Four out of ten requesters (R4, R8, R9, R10) proactively mentioned privacy and security considerations, but also stated that privacy and security were secondary concerns [23] — they cared more about *"getting the job done"* (R4). Other requesters expressed concerns about privacy and security-related risks when asking strangers on the Internet for help, especially if the task context — i.e., the action that the CAPTCHA was authenticating — was a monetary transaction or signing up for a new account. Although they felt that volunteers on the Internet might have good intentions to help them, some PVIs still had some concerns — R8, for example, said *"you can never be too careful."*

Requesters' privacy and security concerns varied based on whether they were sourcing help from friends or strangers. When sourcing help from strangers, requesters were most concerned about privacy and security in financial task contexts. They also expressed concerns about what volunteer strangers could access on their computers. When sourcing help from friends, requesters had more concerns about whether their helper could see private information such as browsing history. We asked requesters about concerns they had about using friendsourcing to solve online tasks more generally, where cropping exact section out of the complex screen contents may not always be successful. They expressed a common preference that they do not want their friends to see the whole screenshot and know which website they are visiting (R8: *"I don't want my friends to know where I am looking, but I don't care if strangers see it. They don't know me anyway"*). In contrast, requesters were unconcerned if crowd workers could access their browsing histories. In short, help source appears to affect requesters' privacy/security concerns for assistive transfer systems.

Helpers' privacy and security perceptions of the tool From the helpers' perspective, when we asked them to imagine they were helping PVIs that they did not know personally, some helpers (H3, H6) mentioned that they would not trust some SMS messages containing a link from an app on which they were not pre-registered — as they might be phishing links. They suggested that the tool should also provide a channel for helpers to register beforehand (like BeMyEyes) even though they do not need to install the extension to help PVIs. A registration process would give helpers some confidence and trust in the tool when receiving messages.

5.3 RQ3: Factors Affecting Preference on Source of Help

We found that participants' preferences for sourcing help from close social connections (friendsourcing) or strangers (crowdsourcing) varied based on four factors:

- **Factor 1: Page content**: What content is on the webpage in which the CAPTCHA is embedded? Is there sensitive information?
- **Factor 2: Use case**: What is the PVI doing? What task is being authenticated by the CAPTCHA? Is the action sensitive or security-related?
- **Factor 3: Helper availability**: Is it a good time to ask for help?
- **Factor 4: Impact on social relationship**: Does the PVI feel like they are burdening their helpers?

Most participants (7 out of 10 requesters) preferred sourcing help from their friends/family for small favors such as solving CAPTCHAs. They mentioned that asking friends for help would make them feel more secure than asking strangers, and that requesting small favors such as solving a CAPTCHA would not bother friends much (e.g., R2: *"Yes I would rather ask friends since it's very simple for them,"* Factor 4). We further discuss how assistive transfer systems might impact social relationships in Section 5.5. Participants also expressed a preference for friendsourcing when the use-case was security sensitive, such as a financial transaction (e.g., R4: *"I feel a little uncomfortable when sending this [a CAPTCHA embedded in a money transfer use case] to strangers.'* Factor 2). The three other requesters (R7, R8, R10) who preferred crowdsourcing expressed concerns about bothering their friends (e.g., R7: *"I don't want to interrupt my friend when it's midnight"* Factor 3) and privacy issues (e.g., R8: *"I don't want my friends to know where I am looking"* Factor 1) as discussed in the last subsection. We also found that participants would like to have the option of choosing between friend- and crowdsourcing on a case-by-case basis.

5.4 RQ4: Compensation Preferences

Three out of seven requesters who preferred friendsourcing mentioned saving money as a key rationale for their preference — CAPTCHA tasks are small and simple enough for friends. In comparison, four out of seven requesters still preferred to compensate their friends for helping them, even for small favors such as helping with solving a CAPTCHA.

We also found that nearly every requester (9 out of 10), even the ones who preferred to receive help for free, would prefer to compensate their helper with non-monetary rewards such as *"a cup of coffee."* They believed paying their friends a small amount of money would make them *"feel weird,"* or that it would be an *"insult."* R2 mentioned that a subscription service would also be acceptable for both requesters and helpers since a routine and fixed payment would cause less embarrassment between requesters and helpers.

For the three out of 10 requesters who preferred crowdsourcing, all preferred to use a paid service rather than a free tool. Some requesters (R8, R9) were already using paid remote assistance services to help them with any technical issues they encounter while using computers. These requesters expressed that they would trust the crowd workers more when they paid for the service as the helpers would be *"trained or professional workers."* (R8)

From the helpers' standpoint, most preferred to be compensated with a non-monetary award. For solving a small task like Google reCAPTCHA, they would also settle for completely voluntary work with no payments or rewards. We also asked whether helpers would be willing to help with strangers' requests — most mentioned that they would be willing, at least in theory. Only one helper (H3) mentioned that helping strangers might be overwhelming because *"you need to pick up random messages."* Most helpers mentioned that they would not feel bothered by a small number of PVIs with a small number of requests, e.g., 3-5 friends requesting fewer than 5 times per day. In terms of compensation, most helpers would not expect getting paid much or getting paid at all because *"solving a CAPTCHA only takes seconds"* (H8).

5.5 RQ5: Impact on Social Relationships

Surprisingly, most participants (5 out 10 requesters and 7 out of 8 helpers) thought that an assistive transfer system like WEBALLY would bring requester and helper closer in their relationship (see Figure 4). H3 believed this tool could give them *"more contact opportunities,"* and there is *"value of knowing that I helped my sister."* Some participants thought the tool would not change their relationship since they were already friends, and there were not *"many additional interactions"* (R5). Also, R5 pointed out WEBALLY's social value in raising sighted people's awareness that PVIs need help in these everyday tasks.

We also asked questions about request boundaries and limits for helpers. Most helpers indicated a number between 5 to 8 times a day that they would feel comfortable solving tasks for friends or strangers who need help. If incoming requests exceeded this number, helpers expressed that they would feel bothered; if these requests came from a friend, it would potentially negatively impact their relationship. In a pre-study interview, we asked PVIs how many times they needed help solving CAPTCHAs in their weekly web use — the reported number was less than 10 times weekly, well within the helpers' reported boundaries. While these numbers are specualtive and would need to be validated in a field deployment, these results suggest that friendsourcing may be a viable option for assistive transfer systems that help PVIs solicit just-in-time help for reCAPTCHA tasks.

6 Design Implications

Based on our study findings, we synthesized a number of design implications for designing transferable CAPTCHAs and assistive transfer systems, more broadly.

6.1 CAPTCHAs

Consider Making Inaccessible Tasks Transferable Ideally, CAPTCHAs would simply be more accessible, eliminating the need for assistive transfer systems like WEBALLY altogether. However, given the known and longstanding accessibility issues with CAPTCHAs and their alternatives, we suspect that sweeping changes to improve the accessibility of security challenges will be slow in coming. In the interim, CAPTCHA designers might consider making it easier for assistive transfer systems to work. For example, WEBALLY would likely be too slow unless a friend or helper was on standby. CAPTCHA designers might slightly increase the time allowed for a CAPTCHA to be solved if a transfer request is initiated, for example.

The Interplay between Security and Accessibility CAPTCHAs were originally designed to distinguish humans from online bots. The specific ways of making CAPTCHAs more accessible for PVIs might have security implications — for example, if bots could pose as PVIs in order to trigger these transfer requests. This issue is more prominent in crowdsourcing than in friendsourcing contexts — presumably, friends would need to be pre-registered and only accept requests from those they personally know. In a crowdsourcing context, fees associated with the service may discourage bots from utilizing such requests. For voluntary crowdsourcing services, the onus should be on the service that facilitates such task transference in doing due diligence (e.g., only registered users can send requests with a time-based limit).

6.2 Assistive Transfer Systems

Our paper focused on exploring the design space of assistive transfer systems — i.e., a system that assists PVIs by soliciting just-in-time help from friends or helpers — for Google reCAPTCHA. However, our findings offer broader implications for the design of assistive transfer systems.

Complementing, not replacing, existing workflows Given a choice between a CAPTCHA challenge they could solve themselves and a perfect assistive transfer system, PVIs would likely choose the former. Thus, assistive technologies that support independent use of computing devices should be the ultimate goal. However, the modern web is a far cry from being fully accessible for PVIs. At least in the short-term, our research suggests that there is value in allowing accessibility hurdles to be transferred to pre-registered friends and crowdworkers so that PVIs can have a "last resort" option when they have exhausted options to overcome the hurdle independently. While still interdependent, the use of online assistive transfer systems like WEBALLY could remove the need for the helper to be physically present – often how PVIs obtain help from trusted allies – for everyday challenges. We note, however, that interdependence is not inherently bad and could also open new design possibilities for assistive technologies that empower PVIs [11].

Reducing Latency with a Helper List and Speculative Recruitment Most of the PVIs who participated in our study reported having more than one close friend who was available to help. An active friend list would help increase the chance of tasks getting picked up and solved within the time limit. An assistive transfer system should distribute load across many willing friends and/or helpers, and iterate through the list if there is a delay in response. One could also imagine the use of speculative and/or proactive recruiting, à la Bernstein et al.'s Crowds in Two Seconds approach [12]. When a PVI navigates to a website in which a CAPTCHA request is likely to occur, an assistive transfer system might pre-emptively request a friend or helper to be on standby. Of course, care will need to be taken to ensure the PVIs' privacy preferences are respected – they should be offered an informed choice. Another possible option is that helpers can indicate their availability status (e.g., "free," "busy"). Perhaps assistive transfer workflows could also be integrated with social networking websites and instant messaging apps, eliminating the need for a separate setup and effectively using PVIs' existing social connections.

Compensation preferences Our findings suggest that PVIs and their friends, alike, prefer non-monetary compensation for assistive transfer systems like WEBALLY, mirroring Zhu et al.'s findings in prior work [62]. Thus, in practice, designers should consider creative alternatives to payment, e.g., a small gift as a token of appreciation. In crowdsourcing contexts, PVIs seemed to prefer a paid subscription service over transactional micro-payments.

New Opportunities for Social Interactions Contrary to our initial expectations, some of our participants mentioned that occasionally sourcing help from friends to overcome accessibility hurdles would be a good excuse to catch up with that friend. Moreover, PVIs' friends echoed this sentiment, stating that they felt good in the knowledge that they helped their friend. While a field study is necessary to validate this effect in practice, this result suggests that assistive transfer systems have the potential to help maintain or even improve social relationships between PVIs and the friends from whom they source help. Such systems might offer new opportunities for meaningful social interactions.

7 Limitations and Future Work

First, our design probe was only limited to the Google reCAPTCHA challenge and was implemented as a Chrome browser extension. The present research did not explicitly cover other CAPTCHAs, transferable tasks or browsers. However, we believe that the Google reCAPTCHA is one of the most popular CAPTCHAs, which present a common challenge for PVI users. Our tool implementation could also be extended to explore other CAPTCHAs and browsers.

Second, we only asked the participant pairs to solve CAPTCHAs using our tool rather than testing additional ways to solve Google reCAPTCHAs. This was mainly because (1) our transferable approach was meant to complement rather than replace existing solutions, (2) our study goal was to explore the transferable task design space, and (3) we felt that asking PVI participants to directly solve CAPTCHAs in addition to using our tool could be exhausting for them. We also avoided direct comparisons with crowdsourcing assistive tools like BeMyEyes because the afforded functionality is significantly different as solving CAPTCHAs requires more than descriptive guidance.

Third, to explore different design configurations of transferable CAPTCHAs, we asked PVI requesters and sighted helpers to imagine that they are in hypothetical scenarios where (1) the helper is an ally (family members or friends) or a stranger/crowd worker, and (2) the helper gets paid or not for solving the CAPTCHA. However, this role-play still has limitations in ecological validity because for instance, it did not capture ally's availability (e.g., in practice, they might not be available at the time when PVI users need help). A field trial can better represent the realities but it is less suitable for our study goal of exploring the design space of transferable tasks rather than testing the effectiveness of a final system. We also considered having a PVI participant's ally participant serve as another PVI participant's stranger/crowd worker. However, we were not able to recruit/schedule two pairs of participants to do the study at the same time due to people's different schedules. In fact, scheduling a pair of PVI participant and the ally participant to conduct the study together was already very challenging.

Another limitation worth mentioning is that Google reCAPTCHAs will be passed if users are already logged in their Google accounts. However, if the users choose to remain private (e.g., use incognito mode), they will face CAPTCHAs much more frequently. We recognize that PVIs might be in a position where they need to make extra effort to maintain their privacy, and we provide WEBALLY as an additional option when they face such inaccessible challenges.

WEBALLY's current implementation may open the door to a few security risks. For example, since requests are sent to helpers as URLs via SMS, one can imagine a new vector for phishing unsuspecting helpers. Malicious requesters might also use WebAlly to circumvent CAPTCHAs for free. A field-ready implementation could mitigate these risks by requiring requester registration and authentication, allowing helpers to whitelist from whom they can receive requests, and imposing daily request quotas to avoid abusive use.

Finally, the assistive transfer system approach requires interdependence between PVIs and their allies, which might affect PVIs' perceived independence in doing daily tasks. It is important to note, however, that even without WEBALLY, PVIs often ask for help to bypass visual CAPTCHAs. An alternative direction is designing new mechanisms that replace visual CAPTCHAs, for instance, better audio CAPTCHAs that do not require visual abilities. Since visual CAPTCHAs are still the most common type of CAPTCHAs, replacing them in practice will take time and require the creation of new standards. In the meanwhile, assistive transfer systems like WEBALLY can help improve web accessibility for PVIs more immediately.

While we focused on task-based CAPTCHAs in this work, the assistive transfer system approach can also be further explored to support other online tasks, for instance, helping PVI users screen images before they share them on social media to limit potential privacy leakage.

8 Conclusion

To help PVIs overcome task-based visual CAPTCHAs that frustrate and encumber their daily web use, we designed and implemented a proof-of-concept assistive transfer system — WEBALLY — that allows PVIs to source just-in-time, direct help from friends or trained crowd workers. Through an exploratory lab study with recruited PVIs and helpers, we found that both PVIs and helpers had a generally positive impression towards WEBALLY, finding it to be a useful alternative to other accessibility solutions (e.g., audio CAPTCHAs). Participants also found that WEBALLY offered sufficient mitigation to protect PVIs' privacy and security in enabling limited remote control for task transfer. We also discovered several factors that may affect PVI participants' perception towards using WEBALLY, such as the type of website in which the CAPTCHA is embedded, helper availability, and the potential impact such a system might have on a PVI and their helpers' social relationships. Helpers, too, had varied preferences in terms of how frequently they would their help solicited and how they would want to be compensated for their effort.

In conclusion, assistive transfer systems like WEBALLY could serve as a preferred "last resort" alternative for PVIs when they cannot solve reCAPTCHA tasks independently. However, future work is needed to ensure timely recruitment of help (e.g., through proactive and speculative recruitment) and to establish compensation structures with which both PVIs and helpers feel comfortable. More broadly, we foresee assistive transfer systems as a promising new class of assistive technologies that can empower PVIs to overcome web accessibility hurdles related to security and privacy.

9 Acknowledgement

We thank our participants for overcoming difficulties to schedule meetings with several attendees and their insightful and articulated feedback, especially the help from Aditi Shah. We also thank the anonymous reviewers for their thoughtful comments and suggestions. This work was supported in part by the National Science Foundation (NSF Grant CNS-1652497).

References

[1] Dina A Abdrabo, Tarek Gaber, and M. Wahied. Assistive technology solution for blind users based on friendsourcing. In *The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Beni Suef, Egypt*, pages 413–422. Springer, 2016.

[2] Tousif Ahmed, Roberto Hoyle, Kay Connelly, David Crandall, and Apu Kapadia. Privacy concerns and behaviors of people with visual impairments. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3523–3532, 2015.

[3] Tousif Ahmed, Apu Kapadia, Venkatesh Potluri, and Manohar Swaminathan. Up to a limit? privacy concerns of bystanders and their willingness to share additional information with visually impaired users of assistive technologies. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3):1–27, 2018.

[4] Tousif Ahmed, Patrick Shaffer, Kay Connelly, David Crandall, and Apu Kapadia. Addressing physical safety, security, and privacy for people with visual impairments. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 341–354, 2016.

[5] Taslima Akter, Tousif Ahmed, Apu Kapadia, and Swami Manohar Swaminathan. Privacy considerations of the visually impaired with camera based assistive technologies: Misrepresentation, impropriety, and fairness. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–14, 2020.

[6] Taslima Akter, Bryan Dosono, Tousif Ahmed, Apu Kapadia, and Bryan Semaan. " I am uncomfortable sharing what i can't see": Privacy concerns of the visually impaired with camera based assistive applications. In *29th USENIX Security Symposium (USENIX Security 20)*, 2020.

[7] Anti-CAPTCHA. https://anti-captcha.com/. Accessed: 2021-05-23.

[8] Paul Baecher, Niklas Büscher, Marc Fischlin, and Benjamin Milde. Breaking reCAPTCHA: a holistic approach via shape recognition. In *IFIP International Information Security Conference*, pages 56–67. Springer, 2011.

[9] Daniel Robert Bateman, Erin Brady, David Wilkerson, Eun-Hye Yi, Yamini Karanam, and Christopher M Callahan. Comparing crowdsourcing and friendsourcing: a social media-based feasibility study to support alzheimer disease caregivers. *JMIR research protocols*, 6(4):e56, 2017.

[10] BeMyEyes. https://www.bemyeyes.com/. Accessed: 2021-05-23.

[11] Cynthia L Bennett, Erin Brady, and Stacy M Branham. Interdependence as a frame for assistive technology research and design. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 161–173, 2018.

[12] Michael S Bernstein, Joel Brandt, Robert C Miller, and David R Karger. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 33–42, 2011.

[13] Michael S Bernstein, Desney Tan, Greg Smith, Mary Czerwinski, and Eric Horvitz. Personalization via friendsourcing. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 17(2):1–28, 2008.

[14] Rudy Berton, Ombretta Gaggi, Agnieszka Kolasinska, Claudio Enrico Palazzi, and Giacomo Quadrio. Are captchas preventing robotic intrusion or accessibility for impaired users? In *2020 IEEE 17th Annual Consumer Communications & Networking Conference (CCNC)*, pages 1–6. IEEE, 2020.

[15] Jeffrey P Bigham and Anna C Cavender. Evaluating existing audio captchas and an interface optimized for non-visual use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1829–1838, 2009.

[16] Jeffrey P. Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C. Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, and others. VizWiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342. ACM, 2010.

[17] Kevin Bock, Daven Patel, George Hughey, and Dave Levin. uncaptcha: a low-resource defeat of recaptcha's audio challenge. In *11th USENIX Workshop on Offensive Technologies (WOOT 17)*, 2017.

[18] Richard E Boyatzis. *Transforming qualitative information: Thematic analysis and code development*. sage, 1998.

[19] Erin Brady, Meredith Ringel Morris, and Jeffrey P Bigham. Friendsourcing for the greater good: perceptions of social microvolunteering. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

[20] Anjali Avinash Chandavale, Ashok M Sapkal, and Rajesh M Jalnekar. Algorithm to break visual captcha. In *2009 Second International Conference on Emerging Trends in Engineering & Technology*, pages 258–262. IEEE, 2009.

[21] Geetest CAPTCHA demo. https://www.geetest.com/en/demo. Accessed: 2021-05-23.

[22] Bryan Dosono, Jordan Hayes, and Yang Wang. "I'm Stuck!": A Contextual Inquiry of People with Visual Impairments in Authentication. In *Symposium on Usable Privacy and Security (SOUPS)*, 2015.

[23] Paul Dourish, Rebecca E Grinter, Jessica Delgado De La Flor, and Melissa Joseph. Security in the wild: user strategies for managing security as an everyday, practical problem. *Personal and Ubiquitous Computing*, 8(6):391–401, 2004.

[24] Valerie Fanelle, Sepideh Karimi, Aditi Shah, Bharath Subramanian, and Sauvik Das. Blind and human: Exploring more usable audio captcha designs. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pages 111–125, 2020.

[25] Buster: CAPTCHA Solver for Humans. https://chrome.google.com/webstore/detail/buster-captcha-solver-for/mpbjkejclgfgadiemmefgebjfooflfhl. Accessed: 2021-05-23.

[26] Jennifer LoCascio Gauvreau. Accessibility gotchas with captchas. *Journal of Digital & Social Media Marketing*, 5(4):379–390, 2017.

[27] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. Vizwiz-priv: a dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 939–948, 2019.

[28] Jordan Hayes, Smirity Kaushik, Charlotte Emily Price, and Yang Wang. Cooperative Privacy and Security: Learning from People with Visual Impairments and Their Allies. In *Proceedings of Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, 2019.

[29] Kieran Healy. *Last best gifts: Altruism and the market for human blood and organs*. University of Chicago Press, 2010.

[30] José María Gómez Hidalgo and Gonzalo Alvarez. Captchas: An artificial intelligence application to web security. In *Advances in Computers*, volume 83, pages 109–181. Elsevier, 2011.

[31] J. Holman, J. Lazar, and J. Feng. Investigating the Security-related Challenges of Blind Users on the Web. In Patrick Langdon BSc, CEng John Clarkson MA MIEE, and Peter Robinson MA Ceng, editors, *Designing Inclusive Futures*, pages 129–138. Springer London, 2008.

[32] Mohit Jain, Rohun Tripathi, Ishita Bhansali, and Pratyush Kumar. Automatic generation and evaluation of usable and secure audio recaptcha. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*, pages 355–366, 2019.

[33] Kiranjot Kaur and Sunny Behal. Captcha and its techniques: a review. *International Journal of Computer Science and Information Technologies*, 5(5):6341–6344, 2014.

[34] Sushama Kulkarni and Hanumant Fadewar. Audio captcha techniques: A review. In *Proceedings of the Second International Conference on Computational Intelligence and Informatics*, pages 359–368. Springer, 2018.

[35] Jonathan Lazar, Jinjuan Feng, Tim Brooks, Genna Melamed, Brian Wentz, Jon Holman, Abiodun Olalere, and Nnanna Ekedebe. The soundsright captcha: an improved approach to audio human interaction proofs for blind users. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 2267–2276, 2012.

[36] Il'dar Nailevich Manashev. Breaking google recaptcha v2. *Prikladnaya Diskretnaya Matematika. Supplement*, (11):99–101, 2018.

[37] Lourdes Moreno, María González, and Paloma Martínez. Captcha and accessibility. *Is this the best we can do*, 2014.

[38] Greg Mori and Jitendra Malik. Recognizing objects in adversarial clutter: Breaking a visual captcha. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 1, pages I–I. IEEE, 2003.

[39] Meredith Ringel Morris, Kori Inkpen, and Gina Venolia. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 662–673, 2014.

[40] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. What do people ask their social networks, and why? a survey study of status message q&a behavior. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1739–1748, 2010.

[41] Donatella Pascolini and Silvio Paolo Mariotti. Global estimates of visual impairment: 2010. *The British Journal of Ophthalmology*, 96(5):614–618, May 2012.

[42] Eyecoming Platform. https://www.eyecoming.com/english/. Accessed: 2021-05-23.

[43] Frances L Rapport and CJ Maggs. Titmuss and the gift relationship: altruism revisited. *Journal of advanced nursing*, 40(5):495–503, 2002.

[44] JAWS Screen Reader. https://www.freedomscientific.com/products/software/jaws/. Accessed: 2021-05-23.

[45] NVDA Screen Reader. https://www.nvaccess.org/. Accessed: 2021-05-23.

[46] Google reCAPTCHA. https://www.google.com/recaptcha/about/. Accessed: 2021-05-23.

[47] Jeffrey M Rzeszotarski and Meredith Ringel Morris. Estimating the social costs of friendsourcing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2735–2744, 2014.

[48] Stuart E Schechter, Rachna Dhamija, Andy Ozment, and Ian Fischer. The emperor's new security indicators. In *2007 IEEE Symposium on Security and Privacy (SP'07)*, pages 51–65. IEEE, 2007.

[49] Sajad Shirali-Shahreza and M Hassan Shirali-Shahreza. Accessibility of captcha methods. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pages 109–110, 2011.

[50] Suphannee Sivakorn, Jason Polakis, and Angelos D Keromytis. I'm not a human: Breaking the google recaptcha. *Black Hat*, pages 1–12, 2016.

[51] Online Survey. Webaim: screen reader user survey #7 results.

[52] Jennifer Tam, Jiri Simsa, David Huggins-Daines, Luis Von Ahn, and Manuel Blum. Improving audio captchas. In *Symposium On Usable Privacy and Security (SOUPS)*, 2008.

[53] Kathleen D Vohs, Nicole L Mead, and Miranda R Goode. The psychological consequences of money. *science*, 314(5802):1154–1156, 2006.

[54] Michael Walzer. *Spheres of justice: A defense of pluralism and equality*. Basic books, 2008.

[55] Yang Wang. Inclusive Security and Privacy. *IEEE Security & Privacy*, 16(4):82–87, August 2018.

[56] Webvisum. https://www.webvisum.com/en/main/download. Accessed: 2021-05-23.

[57] Jacob O Wobbrock, Shaun K Kane, Krzysztof Z Gajos, Susumu Harada, and Jon Froehlich. Ability-based design: Concept, principles and examples. *ACM Transactions on Accessible Computing (TACCESS)*, 3(3):1–27, 2011.

[58] Shaomei Wu and Lada A Adamic. Visually impaired users on an online social network. In *Proceedings of the sigchi conference on human factors in computing systems*, pages 3133–3142, 2014.

[59] Huichuan Xia, Yang Wang, Yun Huang, and Anuj Shah. "Our Privacy Needs to Be Protected at All Costs": Crowd Workers' Privacy Experiences on Amazon Mechanical Turk. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW):113:1–113:22, December 2017.

[60] Jeff Yan and Ahmad Salah El Ahmad. Breaking visual captchas with naive pattern recognition algorithms. In *Twenty-Third Annual Computer Security Applications Conference (ACSAC 2007)*, pages 279–291. IEEE, 2007.

[61] Yuan Zhou, Zesun Yang, Chenxu Wang, and Matthew Boutell. Breaking google recaptcha v2. *Journal of Computing Sciences in Colleges*, 34(1):126–136, 2018.

[62] Haiyi Zhu, Sauvik Das, Yiqun Cao, Shuang Yu, Aniket Kittur, and Robert Kraut. A market in your social network: The effects of extrinsic rewards on friendsourcing and relationships. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 598–609, 2016.

[63] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. Research through design as a method for interaction design research in hci. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 493–502, 2007.

A Appendix

A.1 Pre-Study Interview (Requester Only)

- **I know what a CAPTCHA is and have encountered it in my real-life prior to this study**
  o Yes
  o No

- **Roughly, how many times did you need to solve a CAPTCHA in the past 7 days**

- **Roughly, how many times did you need to solve a CAPTCHA in the past 30 days**

- **Where did you usually encounter the CAPTCHA tasks**

- **I am confident in solving a Google reCAPTCHA vision task on my own (the task that ask users to click the right tile images)**
  o Strongly Agree
  o Agree
  o Neither Agree nor Disagree
  o Disagree
  o Strongly Disagree

- **I am confident in solving a Google reCAPTCHA audio task on my own (the task that ask users to type in words they hear)**
  o Strongly Agree
  o Agree
  o Neither Agree nor Disagree
  o Disagree
  o Strongly Disagree

- **Can you briefly describe why are you confident or not in solving these tasks?)**

- **Please explain your obstacles/challenges when solving a CAPTCHA**

- **When I need help for solving a CAPTCHA, I am confident that someone will always be able to help me in a timely fashion**
  o Strongly Agree
  o Agree
  o Neither Agree nor Disagree
  o Disagree
  o Strongly Disagree

A.2 Post-Study Interview

A.2.1 Requester Interview

- **In general, what do you think about the tool?**

- **What was good or bad? What could be improved?**

- **Do you have any concerns using the tool? Can you explain them briefly?**

- **We saw that you chose [some methods] in solving the CAPTCHA tasks, Could you tell us why you chose the method(s)?**

- **If you have the choices, which option(s) would you choose to solve CAPTCHAs in the future? (Please explain why)**
  o Free friend-sourced tool
  o Free crowd-sourced tool
  o Paid fried-sourced tool
  o Paid crowd-sourced tool
  o I'd not choose any of the above options

- **I am confident in solving a CAPTCHA with free, friend-sourcing using the tool (Please explain**
  o Strongly Agree
  o Agree
  o Neither Agree nor Disagree
  o Disagree
  o Strongly Disagree

- **I am willing to install and use the tool**
  o Strongly Agree
  o Agree
  o Neither Agree nor Disagree
  o Disagree
  o Strongly Disagree

- **I am willing to recommend the tool to others**
  o Strongly Agree
  o Agree
  o Neither Agree nor Disagree
  o Disagree
  o Strongly Disagree

- **The tool will bring me and my helping friend closer**
  o Strongly Agree
  o Agree

o Neither Agree nor Disagree
o Disagree
o Strongly Disagree

- **I have privacy or security concerns when using the friend-sourced tool**

  o Strongly Agree
  o Agree
  o Neither Agree nor Disagree
  o Disagree
  o Strongly Disagree

- **I have privacy or security concerns when using the crowd-sourced tool**

  o Strongly Agree
  o Agree
  o Neither Agree nor Disagree
  o Disagree
  o Strongly Disagree

- **Do you have any other suggestions for improving this tool?**

A.2.2 Helper Interview

- **In general, what do you think about the tool?**
- **What was good or bad? What could be improved?**
- **Do you have any concerns using the tool? Can you explain them briefly?**
- **If you have the choices, which option(s) would you choose to help solve CAPTCHAs in the future? (Please explain why)**

  o Helping your friend/family member without getting paid
  o Helping your friend/family member and getting paid for a small amount
  o Helping strangers without getting paid
  o Helping your strangers and getting paid for a small amount
  o I'd not choose any of the above options

- **I am willing to be prompted and help the requester (Please explain** o Strongly Agree
  o Agree
  o Neither Agree nor Disagree
  o Disagree
  o Strongly Disagree

- **I think this tool will be too time-consuming (Please explain)** o Strongly Agree
  o Agree
  o Neither Agree nor Disagree
  o Disagree
  o Strongly Disagree

- **If requesters request it, I can always respond timely**
  o Strongly Agree
  o Agree
  o Neither Agree nor Disagree
  o Disagree
  o Strongly Disagree

- **The tool will bring me and my helping friend closer**
  o Strongly Agree
  o Agree
  o Neither Agree nor Disagree
  o Disagree
  o Strongly Disagree

- **I have privacy or security concerns when helping friends** o Strongly Agree
  o Agree
  o Neither Agree nor Disagree
  o Disagree
  o Strongly Disagree

- **I have privacy or security concerns when helping strangers** o Strongly Agree
  o Agree
  o Neither Agree nor Disagree
  o Disagree
  o Strongly Disagree

- **Do you have any other suggestions for improving this tool?**

Designing Toxic Content Classification for a Diversity of Perspectives

Deepak Kumar[◁] Patrick Gage Kelley[○] Sunny Consolvo[○] Joshua Mason[†] Elie Bursztein[○]
Zakir Durumeric[◁] Kurt Thomas[○] Michael Bailey[†]
[◁]*Stanford University* [○]*Google* [†]*University of Illinois at Urbana-Champaign*

Abstract

In this work, we demonstrate how existing classifiers for identifying toxic comments online fail to generalize to the diverse concerns of Internet users. We survey 17,280 participants to understand how user expectations for what constitutes toxic content differ across demographics, beliefs, and personal experiences. We find that groups historically at-risk of harassment—such as people who identify as LGBTQ+ or young adults—are more likely to to flag a random comment drawn from Reddit, Twitter, or 4chan as toxic, as are people who have personally experienced harassment in the past. Based on our findings, we show how current one-size-fits-all toxicity classification algorithms, like the Perspective API from Jigsaw, can improve in accuracy by 86% on average through personalized model tuning. Ultimately, we highlight current pitfalls and new design directions that can improve the equity and efficacy of toxic content classifiers for all users.

1 Introduction

Online hate and harassment is a pernicious threat facing 48% of Internet users [52]. In response to this growing challenge, online platforms have developed automated tools to take action against toxic content (e.g., hate speech, threats, identity attacks). Examples include Yahoo's abusive language classifier trained on crowdsourced labels attached to news comments [43], Google Jigsaw's Perspective API, which is trained on Wikipedia moderation verdicts for abuse as well as samples from other online communities [35, 57], and Instagram's recent classifier that detects harassing comments posted as a reply to photos [32].

Although platforms have used these classifiers to address toxic content in direct violation of their policies [41], a variety of content that is not toxic enough to violate policy may still cause harm to Internet users [1]. These "gray areas" stem from the fact that users may disagree about what constitutes toxic content online based on their lived experiences, cultural perspective, political views towards free speech, or access to appropriate context [26, 50]. While prior research has demonstrated that certain groups are more at-risk of experiencing online hate and harassment [45, 52], no study has investigated how users from diverse backgrounds interpret online toxicity or how their views on what content they would like to see online differ. Understanding these nuanced differences is an important first step to designing harassment defenses for diverse Internet users.

In this work, we investigate divergent user interpretations of toxic content and identify whether current classifiers can be tuned to accommodate a diversity of perspectives. At the core of our study, we develop a survey instrument that asks 17,280 participants to rate and label the toxicity of 20 random comments drawn from 107,620 Twitter, Reddit, and 4chan comments. In tandem, we collect demographic data and log participants' previous exposure and experiences with online harassment. Taken together, our survey instrument provides access to a diverse set of perspectives on why people deem certain comments as toxic. We explore this data in three steps: we investigate user ratings of toxic content in aggregate, we identify the factors that result in identical comments receiving divergent ratings, and finally, we demonstrate how modern classifiers can better accommodate differing user perspectives.

Participants frequently disagree on whether comments are toxic. In aggregate, participants labeled 53% of our dataset as "not toxic", 39% as "slightly" or "moderately toxic" and the remaining 8% as "very" or "extremely toxic". However, 85% of comments exhibited some form of disagreement, including whether participants were comfortable seeing the comment on any online platform. Even when participants uniformly agree that a comment is toxic, they disagree about the subcategory the comment belonged to (e.g., a threat versus an insult). As such, a la carte models that isolate individual classes of toxic content—for instance, identity-based attacks [55]—may fail

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

to adequately meet the needs of a user base with diverse perspectives on toxic content.

A variety of factors influence how users perceive toxicity. We find that a participant's personal experience with harassment, whether the participant belongs to an at-risk group frequently targeted by harassment [13,45], and a participant's attitudes towards filtering online discourse all correlate with rating a comment as toxic or not. For example, holding all other factors constant, the odds that a participant rates a comment as toxic increase 1.64 times if they identify as LGBTQ+. Alternatively, these odds decrease by 0.78 times for users who regularly witness others targeted by toxic content, potentially due to desensitization. Combined, no single demographic variable or experience defines how participants interpret toxic content, underscoring the need for diverse raters in data labeling and model construction.

Finally, we investigate how we might leverage current state-of-the-art classifiers to enable diverse user perspectives of toxic content online. We focus on Jigsaw's Perspective API [23] and Instagram's comment nudge [32]. As a baseline, we find for content that Perspective deemed 90% likely to be toxic, only 50% of our participants agreed. Similarly, Instagram's classifier flagged only 27% of comments that a majority of our participants rated as toxic. We propose potential improvements based on *personalized tuning*—finding a threshold for the classifier that is set based on individual responses or in larger demographic groups. These improvements achieve an 86% boost in accuracy per individual and a 22% improvement in accuracy per demographic cohort, highlighting personalized modeling as a future direction in toxicity classification.

We conclude with a discussion of how to overcome the limitations of crowdsourced labeling and one-size-fits-all classification that we identified through our work. To this end, we have shared our results with Jigsaw and have released our labeled dataset[1] to enable other researchers to reproduce our analysis, build new classifiers, and further explore how different individuals perceive toxic behavior online.

2 Background & Related Work

2.1 What is toxic content?

We use the term *toxic content* as an umbrella for identity-based attacks such as racism on social media [2,21,55], bullying in online gaming or replies to posts [36,50], trolling [10], threats of violence, sexual harassment, and more [47, 52]. These attacks represent a subset of abuse stemming from *hate and harassment*, a broader threat that encompasses any activity where an attacker attempts to inflict emotional harm on a target (e.g., stalking, doxxing, sextortion, and intimate partner violence) [11,52]. Unlike spam, phishing, or related abuse classification problems that can rely on expert raters, toxic content is an inherently subjective problem. For the purposes of our study, we focus exclusively on text-based toxic content, but attacks may also extend to images and videos [58].

Previous studies have shown that some demographic cohorts in the United States are more likely to receive and report toxic content than others [12,45]. For example, a survey by Pew found that men were more likely to report experiencing offensive name calling and physical threats, while women were more likely to experience sexual harassment [45]. Beyond gender, Black adults were found to report higher rates of name calling and purposeful embarrassment [45], while people who identify as LGBTQ+ were three times as likely to report offensive name calling, physical threats, and sexual harassment [6, 13]. Similarly detailed demographic studies from various global perspectives are not yet available. In order to ensure that automated detection works for all people, including at-risk groups, we argue that it is critical to first understand how different people perceive toxic content and how perceptions generalize across Internet users.

2.2 Detecting toxic content

Security researchers and practitioners have proposed a multitude of blocklist-based, machine learning, and natural language processing techniques to detect toxic content. The simplest of these approaches rely on manually curated lists of abusive words or users, such as HateBase's corpus of hate speech related terms [27], or BlockTogether's list of abusive Twitter accounts [34]. These provide targeted protections against exact matches of terms or known abusers, but fail to generalize to other types of toxic content, or in the context of blocklists, anonymous posts.

More sophisticated machine learning models include Yahoo's regression model trained on a corpus of roughly 300,000 abusive comments with crowdsourced labels that included hate speech, derogatory messages, and profanity [43]. Using a variety of NLP-based features, they found their classifier could achieve an AUC of 0.90, though domain-specific language and concept drift (e.g., changes in abusive terms) degraded performance over time. Since then, a variety of models have incorporated crowdsourced labels such as Wikipedia moderation decisions [18,57], in-game conversations [4], and social media posts [9,15,17,19,51,54] to varying degrees of success. In another example, Founta et al. leveraged HateBase to build crowdsourced sublabels from participants for abusive tweets, and then characterized a sample of Twitter data [22]. Related approaches have examined how to take a model trained for one community and apply it to a separate community or site to avoid the cost of generating a labeled training set [8]. Finally, several studies have focused on latent annotator bias in datasets [46,56] and also demonstrated that disagreements between raters for social tasks may explain why classifiers excel on benchmarks but suffer in practice [24].

[1] https://data.esrg.stanford.edu/study/toxicity-perspectives

Prominent models deployed at-scale today include Jigsaw's Perspective API, a deep learning classifier for detecting toxic comments which is used by the New York Times, Disqus, and other news sites for moderating toxic comments [23]. Similarly, Instagram recently deployed a model for nudging users away from posting comments that the classifier perceives as harassment due to similar abusive text being reported in the past [32]. We evaluate how these models generalize across users in Section 6.

2.3 Other intervention strategies

While our work focuses on how best to train classifiers to automatically detect toxic content, researchers have also considered a variety of other strategies for moderating toxic content. One example is building mechanisms into online platforms to escalate conflicts to community tribunals who are empowered to remove toxic content and take action against abusive users [40]. Other examples include enabling bystanders to simply report toxic content [16], or providing family and friends with tools to assist in moderating toxic content on behalf of a target [5, 39]. All of these techniques leverage community and context to overcome the limitations of automated classification, but alone may fail to scale to the hundreds of millions of interactions that happen online every day. Additionally, these systems cannot relieve moderators of the emotional burden of reviewing toxic content [42].

2.4 Differentiation from prior work

Prior work in evaluating automated toxicity classifiers has focused on either investigating underlying bias in training data, such as flagging comments with the word "gay" as hateful [14, 18], or shown that classifiers are easily manipulated by substituting "offensive" words while retaining semantic meaning [33]. The focus of our work is to first, understand how perspectives of toxic content change based on individual experiences, and second, evaluate the impact these experiences have on automated toxic content detection (Section 6). Prior work identified certain groups to be at higher risk of online harassment [13, 45], however, no work has shown whether these experiences lead to differences in perception of toxic content online. Closest to this is work by Cowan et al. who investigated perceptions of hate speech against three target groups on college campuses. However, their study is limited in scale ($N < 500$) and not specific to an online context; our work focuses on a broader set of participants, focuses on several categories of toxic content, and is more representative of online discussion. Furthermore, we investigate if implementing a personalized filter—one that better captures the sentiment of participants by their individual experiences—can improve toxicity detection.

Theme raised by participants	**Offensive** N=72 $\kappa = 0.95$	**Hateful** N=74 $\kappa = 0.98$	**Toxic** N=79 $\kappa = 0.9$
Insulting, demeaning, or derogatory	42–44%	55%	58–62%
Identity attack, hate speech, or racist	33–35%	39–41%	33–34%
Profane or obscene	21%	12%	19%
Threatening or intimidating	11%	11%	16%
Not constructive or off-topic	3%	0%	9–11%
None of the above	29–32%	20–22%	19–20%

Table 1: **Interpretation of the Terms: Offensive, Hateful, and Toxic**—We find the term toxic resulted in the broadest interpretation for our rating task.

3 Methods

3.1 Survey instrument

Our survey consisted of three parts: pre-exercise questions about the participant's attitude towards technology and toxic content, an exercise where the participant rated 20 comments from social media and community forums as toxic or not, and finally, demographic and attention check questions. We provide our full survey instrument in the Appendix. Our study was approved by our institution's IRB.

Selecting terminology and comprehension. As a preliminary step, we first determined what terminology to use for our rating task. An inherent challenge here is the ambiguity of the term *toxic content* or *hate and harassment* and a lack of consensus across researchers and industry [44].

In the absence of common best practices, we ran a pilot study with $N = 300$ participants recruited from Mechanical Turk to identify the terminology we should use in our survey instrument. We asked each participant the open ended question: "When you see a post or comment, what do you look for to decide if it's $< x >$?", where x was one of "hateful", "offensive", or "toxic". We recruited $N = 100$ participants per survey variant. We did not use the term "abusive" as not to overload its meaning with other online abuse such as for-profit cybercrime or unsafe content including drugs or self-harm. After filtering for attention checks, we received a total of $N = 225$ responses.

We reviewed each response and identified five emergent themes, detailed in Table 1. Two independent raters coded every response according to these themes, with multiple themes possible per response. Coding achieved an interrater agreement Cohen's kappa $\kappa > 0.9$ for all three variants, indicating strong agreement.[2] We found that participants most often interpreted "offensive" to mean comments that were insulting, profane, or an identity-based attack. Participants even more narrowly construed the term "hateful" to mean comments that

[2] In the event that a rater ascribed multiple themes to a single open ended response, we required both raters to select the same set of themes to constitute agreement.

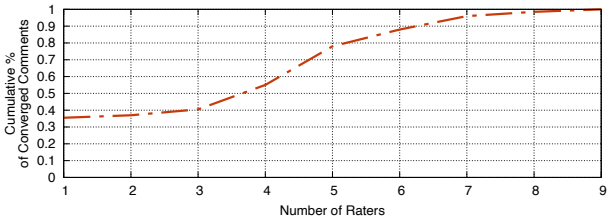


Figure 1: **Toxicity Convergence**—We observe an inflection point where after five participants rate a comment, the benefit of additional perspectives falls off.

involved an identity-related attack or insult. On the other hand, "toxic" encompassed the largest set of themes, where participants also considered whether a comment was constructive or off-topic, and whether a comment was threatening. Based on our findings, we adopted "toxic" as our final survey term to describe our rating task to participants.

Determining the number of ratings per comment. Our survey instrument had to satisfy two competing goals: capturing a diverse enough set of ratings to measure divergence among participants while also maximizing the number of comments rated by participants to produce a meaningfully-sized evaluation corpus. In order to identify how many ratings we should solicit per comment, we ran a pilot survey where 100 participants rated a fixed set of 200 manually curated comments. Each comment was rated by 10 unique participants. Participants selected their rating on a five-point Likert scale ranging from "Not at all toxic" to "Extremely toxic".

We then measured how quickly each comment's ratings converged to its average toxicity score. In this context, we define the average toxicity score to be the fraction of participants that labeled a comment as "Moderately toxic" or greater per comment (the top three ratings of our Likert scale). The global average is the average across all raters for each comment. We then measured the number of participants required for the running average rating to fall within 10% of the global average toxicity score per comment.

Figure 1 shows a CDF of the number of ratings required for convergence for our pilot data. With only 2 ratings, 37% of comments had converged to their final distribution. However, with five ratings, we found 78% of the comments had converged to their final distribution with each incremental participant adding only marginal improvements toward the global average. As we needed to balance soliciting as many ratings as possible per comment with the cost of doing so, we selected five participants to rate each comment for this study.

3.2 Sourcing potentially toxic content

We sourced an initial corpus of 549,058 comments from Twitter, Reddit, and 4chan for our study. We selected these platforms as they represent a diverse cross-section of Internet users, are conversation driven, and contain varying degrees of toxic behavior [3, 7, 28]. All data was collected between December 2019 and August 2020. While our dataset does not capture all types of conversations—such as private discussions via messaging apps or "walled gardens" like Facebook—our collection strategy avoids privacy constraints that would otherwise prevent sharing content with random participants on crowdsourcing platforms.

Stride	Aggregate Rating	% Agreement	% Final Dataset
0.0—0.1	Not toxic	90%	5%
0.1—0.2	Not toxic	81.8%	5%
0.2—0.3	Not toxic	80%	5%
0.3—0.4	Not toxic	76.4%	10%
0.4—0.5	Not toxic	71.4%	10%
0.5—0.6	Not toxic	65.2%	15%
0.6—0.7	Not toxic	68.3%	15%
0.7—0.8	Toxic	65.2%	20%
0.8—0.9	Toxic	76.4%	10%
0.9—1.0	Toxic	80%	5%

Table 2: **Interrater Agreement per Stride**—Although raters agree broadly for comments with either low or high toxicity scores, raters show minimal agreement when a comment is scored between 0.5—0.8. As such, we oversample these ranges for our dataset.

Given the class imbalance inherent to each site, where benign content far outweighs toxic content (with the exception of perhaps 4chan), a purely random sampling approach would be prohibitively expensive to gather crowdsourced labels for a sufficiently large volume of toxic content. Instead, we leveraged the Perspective API `TOXICITY` model (discussed in detail in Section 2) to build a stratified sample of potentially toxic content.[3] The API takes as input a sample of text and returns a score between 0 and 1, describing the likelihood that an audience would perceive the text to be toxic.

In order to identify which score ranges correlated with the largest rating disagreement among participants, we ran a pilot survey where 200 participants rated 800 comments, with 80 comments sourced from each 0.1-stride between 0 and 1. For example, we selected 80 comments with a toxicity score of 0—0.1, 80 comments with a score of 0.1—0.2, and so on. Five independent participants rated each individual comment. We then measured the interrater agreement for each stride as shown in Table 2. We found that participants broadly agreed on comments that had a `TOXICITY` score of < 0.3 or > 0.9, with the least agreement when a comment had a score of between of 0.5 and 0.6 and between 0.7 and 0.8. A comment with a score of 0.5 might look like:

> "I'm so sick of this mess. The Dems are not good because the Repubs are bad. The Repubs are not good when the Dems are bad. The enemy of your enemy can still be your enemy. #BothPartiesSuck"

[3] Instagram does not provide a public API, thus we did not consider it when building our dataset.

Table 2 shows the final distribution of comments we include per stride. Our dataset preferentially includes comments with lower interrater agreement, however, we note that at least 5% of comments are sampled from each API stride. Our data distribution by source is 67% Twitter comments, 15% Reddit comments, and 18% 4chan comments. We note our final dataset contains at least 16,000 comments per platform. Our sampling skews towards Twitter as we wanted to guarantee a fixed ratio of comments per stride while maintaining a large corpus (N $>$ 100,000) but were limited by fraction of comments available in each stride from 4chan and Reddit.

3.3 Recruitment and validation

We recruited participants for our final survey through Amazon Mechanical Turk to "Participate in a survey about content online". Previous studies have validated the use of Mechanical Turk in security and privacy contexts [48]. Given the scale of this work, we needed to balance overall cost, fair compensation, and the goal of attracting a large and diverse sample of workers across MTurk. After piloting, we decided to pay $1 for completion. Participants took a median of 13 minutes to complete the task. We only recruited participants with at least a 95% approval rating [49] and restricted participants to residents of the United States. All participants were over the age of 18. As our survey instrument collects potentially sensitive demographic information (gender, sexual orientation, race, and more), we provided an option to decline every demographic question. As mentioned previously, our survey was approved by our IRB.

In order to validate a participant's responses, we relied on an attention check question at the end of the survey that asked participants to recall what term we had used throughout the survey (i.e., toxic). Additionally, we included an open ended question asking participants to describe how they define toxic content (akin to our pilot) and set a manually identified threshold on this response. We solicited new participants until we reached our $n = 5$ threshold per comment. Our final dataset consists of 17,280 participants and 107,620 rated comments.

Table 3 outlines the demographic distribution of our participant pool. Participants were evenly split across men and women, with a median age range of 25–34. Most participants identified as White, non-Hispanic (71%), and did not identify as a member of the LGBTQ+ community (81%). Attitudes towards religion were mixed with most participants either deeming religion not important (32%) or very important (31%). Political attitudes were mixed across Liberal, Independent, and Conservative participants. Our participants also split evenly between parents and non-parents. Our sample does not perfectly align to the US Census demographics for all demographic cohorts [53]. However, our modeling results in Section 5 control per demographic cohort and will stay consistent even if some cohorts are over or under sampled. Overall, our recruitment provided access to a variety of groups that historically are more likely to be the targets of toxic content. Our dataset is available at https://data.esrg.stanford.edu/study/toxicity-perspectives.

Demographic	Cohort	% Respondents
Gender	Male	46%
	Female	52%
	Nonbinary	1%
Age	18 – 24	12%
	25 – 34	40%
	35 – 44	25%
	45 – 54	13%
	55 – 64	7%
	65+	3%
Race & Ethnicity	Non-minority	71%
	Minority	29%
LGBTQ+ status	Not LGBTQ+	81%
	LGBTQ+	16%
Religion importance	Not important	32%
	Not too important	12%
	Somewhat important	23%
	Very important	31%
Political attitude	Liberal	40%
	Independent	27%
	Conservative	27%
Parent	Yes	52%
	No	47%

Table 3: **Demographics of Respondents**—Our recruitment strategy provided access to a diverse set of raters, including members of communities that are historically at-risk. Not all percentages sum to 100% due to some participants declining to provide demographic information.

3.4 Ethical considerations

Given that our experiments expose participants to potentially toxic content, on the Mechanical Turk description screen we included an initial warning that described our rating task and the potential harms that might arise from participating. We stated:

> Risks related to this research include feeling targeted or potentially hurt by viewing potentially toxic comments and recalling negative experiences in the past regarding your personal experience with toxic comments online.

At this point, participants could choose to accept the rating task or simply move on without any exposure. After accepting the task, participants consented to a longer agreement, that again reminded participants that they would be exposed to toxic content multiple times. Additionally, our stratified sampling approach avoided most egregious toxic content as detected by existing automated classifiers, where there was unlikely to be any disagreement. This is in line with multiple prior studies that rely on crowdsourcing for toxic content

judgements [4, 18, 35]. Overall, participants voluntarily saw a small number of potentially toxic comments in a short session, most of which were rated to be only moderately toxic and which are most in line with conversations that broadly occur on the Internet.

4 Toxicity, Filtering, and Removal Decisions

We examine how often participants deem a comment toxic and the frequency that participants disagreed in rating the severity of toxicity per comment. Additionally, we explore what classes of toxic content (e.g., sexual harassment, profanity) participants were most aligned in recognizing and ultimately their personal beliefs of whether such content should be allowed online.

4.1 Overall perceived comment toxicity

Each comment in our dataset includes five independent toxicity ratings drawn from a Likert scale ranging from "Not at all toxic" to "Extremely toxic." We considered two strategies for aggregating these ratings into an overall non-binary toxicity score per comment. The first, *max rating*, selects the maximum toxicity rating across all participants for a comment. The second, *median rating*, selects the median rating across all participants. Figure 2 shows the distribution of toxicity scores via these two metrics. Although the maximum rating achieves a relatively even distribution across our Likert scale, this often results from a single outlier among the five raters inflating the measured toxicity, compared to the median rating. As such, we opted for the median rating and use it throughout this work unless otherwise noted.

Overall, 53% of comments in our dataset have a median rating of "Not at all toxic", while only 1% of comments have a median rating of "Extremely toxic". An example of an "Extremely toxic" comment from 4chan was:

> "They're like the polar opposite of limp-wristed smug douchebag homo's [sic] and liberals who buy a Prius and think they're better than everyone else".

Comments that were rated either "Slightly toxic" or "Moderately toxic" make up 39.1% of our labeled dataset. Examples from this category include a Reddit comment like "Kids with hoodies are going to be our future criminals," and 4chan comments like "Women can't be responsible for hiring people. It is foolish to entrust hormonal women to the most important part of the company." These examples show how even mildly toxic comments contain racism and sexism—higher ratings merely represent the perceived intensity of the harassment involved.

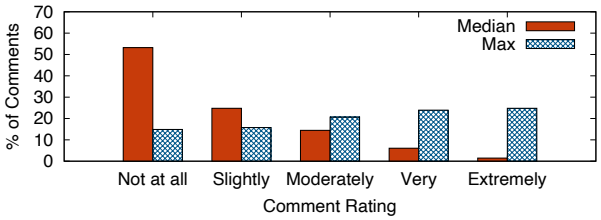


Figure 2: **Median and Maximum Toxicity Per Comment**–The median toxicity rating per comment better captures the central tendency of raters compared to the max rating per comment.

4.2 Subcategories of toxic comments

When participants rated a comment as "Slightly toxic" or higher, they also provided an explanatory label for the toxic behaviors present in the comment. Categories included "Insult," "Identity attack," "Profanity," "Threat," and "Sexual harassment." Participants could select multiple such labels per comment. As an example, the following comment was labeled as both a threat and an identity attack: "I think all you Deep State participants have done Enough [sic] damage to this country.....go jump off the Brooklyn Bridge!". We derived these labels from the themes surfaced by participants in our pilot study, adding "sexual harassment" as an additional theme and removing "off topic" due to the lack of context provided to participants (see Section 3). We refer readers to the Appendix for the detailed instructions that we provided to participants on how to differentiate these categories.

We present a breakdown of the perceived classes of toxic comments in our dataset in Table 4. Each column represents the fraction of comments rated at each toxicity level that fell into each subcategory. Overall, insults are the most common type of toxic comment (67%), followed by profanity (52%), and identity attacks (51%). This is not necessarily an indication that these are the most common toxic behaviors for sites in our sample, but rather these are the toxic behaviors that raters identified. Participants also perceive different sublabels as more or less toxic. For example, 85% of "Extremely toxic" comments involve an identity attack, whereas the same is true for only 57% of comments rated "Slightly toxic" or lower. We also investigate the reverse—which is the fraction of comments in each sublabel that fall into each toxicity level, and find that participants perceive threats and sexual harassment as "Extremely toxic" (3.3%, 3.7% of comments respectively) at a higher rate than identity attacks (2.9%), profanity (2.6%), and insults (2.3%).

4.3 Frequency and intensity of disagreement

While our overall score provides guidance on whether a plurality of participants view a comment as toxic or not, in practice we are interested in how often participants disagree and why. For example, of all comments with a median toxicity of "Not

Category	Overall	Slightly Toxic	Moderately Toxic	Very Toxic	Extremely Toxic
Insult	67%	76%	85%	89%	89%
Profanity	52%	59%	69%	74%	78%
Identity attack	51%	57%	70%	79%	85%
Threat	31%	30%	44%	54%	59%
Sexual harassment	18%	18%	27%	34%	39%

Table 4: **Categories of Toxic Content Recognized by Participants**–Participants were most likely to perceive content as insulting or containing an identity attack, whereas sexual harassment and threats of violence or rape were less frequent.

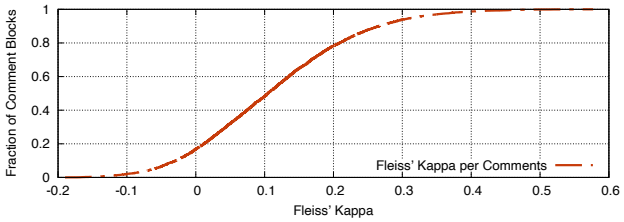


Figure 3: **Interrater Agreement for Subcategory Selection**–Agreement between subcategories of raters is low, with a median interrater agreement score κ of 0.10. The highest agreement between raters only reached 0.57 (moderate agreement), highlighting the difference between rater definitions of subcategories of toxic content.

at all toxic", only 28% have uniform agreement among all five raters. In order to measure diverging perspectives, we calculated the variance of toxicity ratings for each comment. To do this, we treated each rating as an ordinal value between 0 and 4. A variance of 0 indicates perfect agreement for a comment. The maximum variance of 4.8 indicates two competing groups (e.g., two "Extremely toxic," three "Not at all toxic"). We opted for variance over other multi-rater agreement metrics like Krippendorf's alpha or Intra Class Correlation as we are interested in disagreement on individual comments, not between raters.

Only 15% of comments have a variance of 0, indicating all participants rated the comment identically. In aggregate, the median variance of all rated comments is 0.8. However, the spread of scores for comments rated as at least "Slightly toxic" is larger, with a median variance of 1.3 per comment. As an example, the comment from Twitter:

> "At least REDACTED served, unlike you, a weirdo making memes online all day like a little lunatic."

had a variance of 1.3, with two raters finding the comment "Very toxic", one rater finding the comment "Moderately toxic", one finding the comment "Slightly toxic", and one rater not finding the comment toxic at all. In contrast, 7.5% of comments have a variance of 3.0 or greater, indicating widespread disagreement. For example, the comment from Twitter:

> "So you don't want money.... Just free college, loan forgiveness, and (and I'm not sure how this is relevant) healthcare for veterans? I presume you believe only blacks were slaves? Also, your last sentence implies you believe all blacks were slaves..."

had a variance of 3.2. Only 0.03% of comments have a variance of 4.8, which is the maximum amount.

Even when participants agree that a comment has some degree of toxicity, they may still differ on why they feel a comment is toxic. Of comments that participants uniformly deemed toxic, just 0.4% had identical categories assigned by all five participants. We quantify the degree of category disagreement across our dataset using Fleiss' Kappa κ. This score assesses how well a fixed number of raters place a subject into one of several nominal categories—in our case, selecting the same set of categories (e.g., sexual harassment, insult) per comment. In order to arrive at an estimate, we first calculated the κ per block of comments[4] and then calculated the global average.

The best group of five raters achieved a $\kappa = 0.57$, indicating only moderate agreement [37]. The median group of raters achieved a $\kappa = 0.10$, indicating low agreement. These findings illustrate that participants are in general, more likely to agree on toxicity ratings than on the justification for their decision.

4.4 Filtering and removal recommendations

Apart from the perceived toxicity of comments, we also asked participants to make a decision for whether they personally would want to see each comment (e.g., personalized filtering), and whether the comment should be allowed online at all (e.g., global filtering). Of comments rated "Slightly toxic" or higher, participants reported they would personally not want to see 37% of comments. We did not observe a strong distinction between personal filtering and global filtering. In the event a participant felt personal filtering was appropriate, they also felt that the comment should not be allowed online generally 70% of the time. In the most extreme case, 30% of participants would *never* remove a comment from an online platform—even for participants that rated at least one comment as "Extremely toxic" (as 10% of that 30% of our participants did). That participants can recognize harassment but decline intervention represents one of the fundamental conflicts between tackling toxic content online and unfettered free speech.

We observe similar, competing perspectives when it comes to who participants feel is the most responsible for addressing

[4]Each set of five participants are guaranteed to rate the same twenty comments in a random order, which enables us to compare kappa values across participants per block of comments.

toxic content online. As part of our pre-exercise questions, we asked participants whether they felt toxic content was a problem and what party was most responsible for addressing toxic posts or comments online. 42% of participants felt toxic content was very frequently or frequently a problem. Another 51% felt it was rarely or occasionally a problem, while 5% felt it was not an issue at all. Additionally, 47% of participants felt the onus of addressing toxic content was on the user who sent the comment, compared to 27% of participants who felt that the hosting platform held the most responsibility. This rift in beliefs—both for toxic content being an issue online, and what party is responsible for solving it—represents a challenge moving forward for tackling harassment online.

5 Competing Perspectives of Toxicity

Given the frequency of disagreement among raters on what constitutes toxic content, we explore potential explanatory variables stemming from a participant's personal experiences, demographics, and opinions on whether toxic content is a societal problem.

5.1 Modeling participant decision making

We treat each rating task per participant as a Bernoulli trial where a rating of "Moderately toxic" or higher indicates the participant found a comment toxic (e.g., a successful event, or 1), and all other ratings as benign (e.g., failure, or 0). We then model the frequency of success across all labeling tasks as a quasi-Binomial distribution $Y_i(n_i, \pi_i, \phi)$ using a logarithmic link function. The model's parameters consist of categorical variables related to a participant's age, gender, political affiliation, religious beliefs, LGBTQ+ affiliation, education, race and ethnicity, and parental status. The model also incorporates whether a participant has previously witnessed toxic content online or personally been the target of toxic content, whether the participant thinks toxic content is an issue, and who is most responsible for addressing toxic content.

Table 5 contains the results of our model. We report the model's weights as the odds that a participant with a specific trait or belief—after holding all other traits constant—will rate a comment randomly drawn from our corpus as toxic. All results noted with an asterisk are statistically significant with $p < 0.01$. While not shown in a table, we repeat the same modeling process to also understand if any factors influence a participant categorizing a toxic comment as any of our five subcategories of toxic content. We report the full parameters of our models in the Appendix. We discuss the results of our full analysis in detail below.

5.2 Influence of personal experiences

Overall, 77% of participants reported having witnessed toxic content while online. This aligns with a prior Pew study of personal experiences with online harassment, which observed

Demographic	Treatment	Reference	Odds
Gender	Female	Male	0.952
	Non-binary	Male	0.707
Age	18-24	35-44	1.238*
	25-34	35-44	1.227*
	45-54	35-44	0.972
	55-64	35-44	0.980
	65+	35-44	0.977
Race & Ethnicity	Minority	Non-minority	1.126*
LGBTQ+	LGBTQ+	Not LGBTQ+	1.644*
Political affiliation	Conservative	Liberal	1.024
	Independent	Liberal	0.901*
Importance of religion	Not too important	Not important	1.216*
	Somewhat important	Not important	1.572*
	Very important	Not important	1.840*
Parent	Is a parent	Not a parent	1.330*
Education	College	High school	1.139*
	Advanced degree	High school	1.365*
Impact of technology on society	Very negative	Neutral	0.803*
	Somewhat negative	Neutral	0.870
	Somewhat positive	Neutral	0.970
	Very positive	Neutral	1.142*
Toxic content a problem?	Rarely	Not a problem	1.030
	Occasionally	Not a problem	0.958
	Frequently	Not a problem	1.029
	Very frequently	Not a problem	1.125*
Party most responsible	Law enforcement	Bystander	1.282*
	Receiver	Bystander	0.716*
	Platform	Bystander	0.706*
	Sender	Bystander	0.619*
Witnessed toxic content	Yes	No	0.780*
Target of toxic content	Yes	No	1.483*

Table 5: **Demographics, Experiences, and Opinions**—We report the change in likelihood that a participant will flag a random comment as toxic, given a specific trait, in terms of odds. All values noted with an asterisk are significant with $p < 0.01$. See Appendix for model weights and exact significance values.

73% of Americans have observed online harassment [45]. Conversely, 29% of participants in our study reported having been the target of toxic content.[5] Both of these experiences exhibit a statistically significant influence on toxicity ratings. Prior personal experience with being the target of toxic content increases the odds of rating new content as toxic by 1.483 times. These participants potentially empathize with others who might be emotionally harmed by toxic content, and as such, take a stronger stance on what behavior constitutes harassment. Conversely, prior experience with witnessing toxic content decreases the odds of rating new content as toxic by 0.780 times. These participants potentially view

[5] Participants answered both of these questions after the labeling task, which means their answers may have been colored by the perceived toxicity, or lack thereof, of the comments they labeled.

new toxic content through a comparative lens, excusing abusive behavior that does not rise to the level of severity the participant previously encountered. Our findings illustrate the importance of understanding the experience of people who have been targets of harassment as well as highlights the risk of desensitization.

5.3 Influence of demographics

Gender. We find no statistically significant differences between the odds that non-binary, female, and male participants rate a comment as toxic. Furthermore, female and male participants have nearly identical rates for identifying each subcategory of toxic content. One exception is that the odds of a male participant identifying a comment as threatening compared to female participants increases by 1.158 times. One potential explanation is that men report higher rates of physical threats and name calling compared to women [45], and may be more sensitive to those categories of toxic content.

Age. We find that young participants in particular are more likely to flag comments as toxic compared to older participants. Specifically, the odds of rating a comment as toxic by people ages 18–34 increases 1.227–1.238 times compared to participants aged 35–44. When comparing people 35–44 and groups of older adults, we find no statistically significant difference between successive age groups. One possibility is that younger participants may be more represented on the sites we sample from, and thus familiar with the slang or style of attacks present. In line with previous studies [13, 45], participants between the ages of 18–34 also experienced online harassment at higher rates (27%–30% versus 20–24%), which may shape their opinion and sensitivity to toxic content.

LGBTQ+. A participant's LGBTQ+ identity plays a strong role in toxicity ratings. Identifying as LGBTQ+ increases the odds of rating a comment as toxic by 1.644 times compared to participants who do not. Furthermore, LGBTQ+ participants were far more likely to assign all subcategories to toxic comments—with threats showing the largest increase in odds (1.865 times). LGBTQ+ participants are a historically at-risk cohort for online harassment [13] and so may be cognizant of toxic behaviors, biases, and language that other participants fail to identify.

Importance of religion. Religion has one of the strongest influences on how participants perceive toxic content. In particular, religion being "Very important" to a participant increases the odds they rate a comment as toxic by 1.840 times. This impact still holds even when a participant reports that religion is "Not too important", where the odds of rating a comment as toxic increase by 1.216 times. Similarly, religious participants were far more likely assign all subcategories to toxic comments—with profanity and threats showing the largest increase in odds (1.604–1.878 times).

Parents. There is a small but statistically significant difference between the perspectives of parents and non-parents. Being a parent increases the odds of rating a toxic as comment by 1.330 times. Being a parent also increased the odds of flagging sexually harassment (1.298 times) and profanity (1.158 times). These differences are potentially influenced by content that parents do not want their children to see online.

Race and Ethnicity. We find that belonging to a racial or ethnic minority plays only a small role in influencing perspectives of toxic content, amounting to an increase in odds of 1.126 times compared to non-minority participants. Previous studies have shown that minorities and non-minorities experience similar rates of online harassment, but that when harassment occurs, people self-report it is more likely a result of their race or ethnicity [45].

Education and political affiliation. Compared to participants with only a high school education, the odds participants with advanced degrees labeled comments as toxic increases 1.365 times, however, we find no similar relationship to those with college degrees but no advanced degrees. Finally, we find that a participant's political affiliation also has a small impact on the odds of identifying toxic content. Notably, identifying as an independent decreases the odds of flagging toxic content online by 0.901 times compared to liberal participants. These variations may stem from the underlying content and discussions present in our dataset.

5.4 Influence of technology beliefs

Finally, we examine how attitudes towards technology and toxic content online influence toxicity ratings. We find that, when participants feel that toxic content is "Very frequently" a problem, the odds they flag content as toxic increases by 1.125 times compared to others who feel toxic content is "Not a problem". Similarly, when participants feel that technology's role in peoples' lives remains "Very positive", the odds they flag content as toxic increases by 1.142 times compared to neutral participants. These participants potentially have a lower threshold for what they deem to be toxic behavior, or feel a greater obligation to address toxic content.

6 Benchmarking Toxicity Classifiers

Given the influence of personal experiences on toxicity ratings, we next analyze how well widely-deployed automated detection systems from Jigsaw and Instagram currently perform in aggregate, per demographic cohort, and per individual.

6.1 Perspective API

Overall performance. As previously discussed, our dataset uses stratified sampling to oversample potentially toxic comments with the highest rates of disagreement among

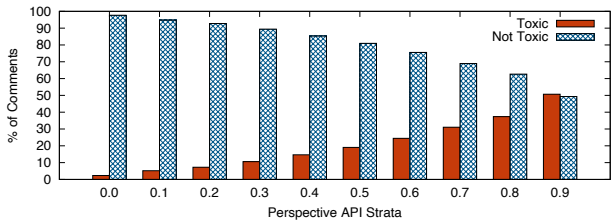


Figure 4: **Toxic/Benign Comment Distribution per Perspective API Stride**—Higher Perspective API scores correlate with a larger fraction of toxic content, however, the fraction of toxic content per stride never exceeds the fraction of benign content.

Stride	% Comments	Accuracy	RMSE
0.0	5.0	0.98	0.12
0.1	5.0	0.95	0.14
0.2	5.0	0.93	0.18
0.3	10.0	0.90	0.24
0.4	10.0	0.85	0.30
0.5	15.0	0.81	0.36
0.6	15.0	0.76	0.42
0.7	20.0	0.50	0.48
0.8	10.0	0.37	0.55
0.9	5.0	0.51	0.55

Table 6: **Accuracy and Root Mean Squared Error per Perspective API Stride**—As Perspective API scores increase, we observe a larger divergence between what participants construe as toxic content versus the classifier. For comments that Perspective rates as 90% likely to be toxic, our participants agreed only 45% of the time.

participants. We omit the vast majority of benign content on Twitter, Reddit, and 4chan that would otherwise be present in a random sample. As such, it is misleading to compare standard performance metrics (e.g., accuracy, precision-recall) across our entire dataset. We control for this bias by considering the accuracy of the Perspective API per *stride* of our sampling. As part of this, we convert every comment's rating distribution into a binary verdict. We treat every comment with a median Likert score of "Moderately toxic" or higher as toxic and all other comments as benign. To compute accuracy, we deem a perspective score of > 0.75 as toxic and all other comments as benign.

Figure 4 shows the fraction of toxic and benign content at each stride of our dataset. The 0.1 stride includes all comments the Perspective API gave a 0–10% likelihood of being toxic, whereas the 0.9 stride includes all comments with a 90-100% likelihood of being toxic. While higher Perspective API scores have monotonically increasing degrees of perceived toxicity, the fraction of toxic content per stride is almost always smaller than fraction of benign content, with the exception of the highest stride, where the labels are roughly equal. Overall, we find only a weak correlation between our participant's Likert ratings and the Perspective API ($r = 0.39$, $p = 0.0$). In line with this, the accuracy for comments in the highest Perspective API stride is only 51%, indicating our participants disagreed with the Perspective rating in 49% of cases. As such, it appears that the Perspective API favors false positives over false negatives. Such a balance is better suited for re-ranking or informing moderation decisions as opposed to outright filtering.

Identifying divergent comments. To provide qualitative examples of divergence between participants and the Perspective API, we isolated comments with the largest intensity of disagreement between the Perspective API and our participants to understand the underlying cause. This is best captured by a root mean square error (RMSE). In particular, we first calculated the average rating across all participants for each comment which we then normalized to a 0–1 scale. We relied on the original 0–1 score of the Perspective API for comparison. We find that error is smallest for lower scores of the Perspective API, indicating both the classifier and our participants agree on what is not toxic. However, error increases as strides increase, with error maximizing at the highest stride ($>= 0.9$) as shown in Table 6. This result is consistent with our observations with accuracy and indicates that the API struggles to match ground truth at high decision thresholds.

We then selected the top 200 comments with the highest error between participants and the Perspective API. Of these, we selected 100 where the Perspective API skewed higher, and 100 where the Perspective API skewed lower. Two independent coders reviewed each comment, came together to form a codebook of high-level themes, and then independently rated every comment before resolving any disagreement. Examples comments where Perspective and our participants disagreed on toxicity ratings can be found in the Appendix, Table 8 and Table 9.

In some cases, the Perspective API is more sensitive to toxic content than our participants. The majority of these comments included some form of profanity that participants did not find toxic (82%). For example, Perspective rates the comment "holy shit, you are perfect!" at 0.83 while every rater found the comment to be not at all toxic. Other sources of disagreement include differing sensitivities to negativity (6%), and scenarios where negative terms convey emphasis (4%). Examples include "Song sucks" and "The stupid tongue! :D Cutest thing" respectively. In effect, the classifier is incorrectly shaping its decision based on negative or profane keywords alone.

In contrast, our participants also occasionally deemed content more toxic than the Perspective API. Identity attacks against minorities, women, and the LGBTQ+ community made up 39% of comments missed by the Perspective API. Another 17% of comments contained political messages that participants deemed toxic. For example, participants rated the comment "YOU'RE JUST LIKE YOUR BOSS AND THE REST OF THE GOP A BUNCH OF!!!!! #CLOWNS" with

a toxicity of 0.75, compared to 0.48 from the Perspective API. Other themes included adult content (11%) and threats of violence or rape (9%). Additionally, despite Perspective API regularly flagging profanity, minor grammatical changes such as the lack of spaces in "nofuckingbody" resulted in score of 0.21, whereas adding spaces results in a score of 0.93. Researchers have abused this sensitivity to minor perturbations in text to construct adversarial examples that evade the Perspective API [25, 29].

Tuning classifiers to personal preferences. Our results indicate that a single definition of toxic content online does not capture the varied experiences and opinions of Internet users. As such, a one-size-fits-all model for abuse detection will likely not be able to capture the toxicity preferences of all participants. Recent work from Google Jigsaw has focused on allowing participants to "Tune" existing APIs to their own personal preferences, simply by adjusting the Perspective API to a specific threshold [38]. However, it is unclear how effective this tuning strategy can be to end-users and where this mechanism may fall short. We investigate the differences in accuracy and precision for the optimal threshold for each individual participant compared to the dataset in aggregate. Although our dataset is not a truly random sample of Internet comments, comparing personal thresholds to the aggregate still provides insight into the effectiveness of personal tuning.

To identify the optimal threshold for all ratings taken in aggregate, we first convert each comment rating into a binary label. A comment rating has a positive label if the participant personally did want to see the comment online, and a negative label if they did not. We then sweep over all Perspective API decision thresholds from 0–1 and identify the *lowest* threshold that maximizes the F1-score, which is the weighted average of the precision and recall. We find that the optimal perspective API threshold for the aggregate dataset ranges from 0.18–0.49, all of which achieve a precision of 0.35 and an accuracy of 0.37.

We perform the same analysis on an individual level, identifying a threshold that maximizes the F1 score for each participant. If a participant did not personally elect to remove any comments they encountered, we set their threshold to the maximum possible value (1.0). Figure 5 shows a distribution of thresholds per individual. For 21.6% of participants, their maximal threshold is 0.0, suggesting that labeling every comment as toxic maximizes both precision and recall. The median threshold is 0.61, resulting in an average precision of 0.6 and an average accuracy of 0.68, an increase in accuracy of 86% compared to a one-size-fits-all classifier. Per this personalized approach, 71.5% of participants saw an improvement in accuracy over the one-size-fits-all optimum accuracy. As such, more research is needed to understand how best to quickly personalize models and how to gather ongoing feedback in order to adjust model thresholds.

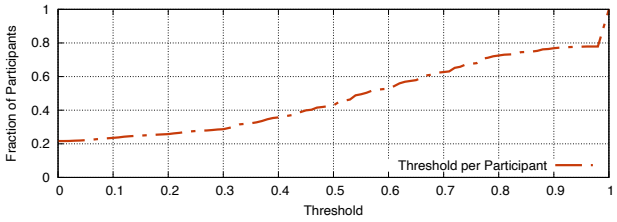


Figure 5: **Optimum Personalized Threshold per Participant**—The threshold that maximizes classifier accuracy per participant is mixed. 21.6% of participants are maximized at a threshold of 0.0, which amounts to labeling every comment as toxic. After tuning to personal thresholds, 71.5% of participants achieved an accuracy greater than the overall classifier.

	Max Precision		Max Accuracy	
Demographic	**Value**	**% Change**	**Value**	**% Change**
Religion	0.40	14.3%	0.41	10.8%
Politics	0.37	5.7%	0.37	0%
Age	0.44	25.7%	0.44	20.6%
Gender	0.39	11.4%	0.40	7.5%
Race	0.36	2.9%	0.36	-2.7%
Parent	0.37	5.7%	0.39	5.4%
LGBTQ+	0.36	2.9%	0.37	0%

Table 7: **Optimum Accuracy and Precision per Demographic**—We show the maximum accuracy and precision when tuning the Perspective API per demographic cohort, as well as the percentage change from the one-size-fits-all model. We find that cohort-based models perform marginally better in some categories, but fall short of performance improvement from personalized models.

Tuning classifiers to demographic preferences. Given differences between demographic cohorts (Section 5), we also investigate the performance benefits for tuning the Perspective model to broad demographic groups. Table 7 shows the maximum precision and accuracy when each independent demographic group is tuned for separately. We find that demographic tuning in aggregate offers a smaller improvement over the aggregate classifier compared to personalized tuning, with only a 0–20.6% increase in accuracy. Age-specific model thresholds provided the best performance gain. These results highlight that even within broad demographic groups, individual experiences and preferences take more importance when making toxicity determinations online. Any cohort-based model would need to account for multiple factors when designed and deployed.

6.2 Instagram nudges

In December 2019, Instagram rolled out a feature that nudges a user if they are about to post a comment similar to those that have been flagged in the past. As a small experiment, we also compare how well the Instagram classifier performs against

our ground truth data. We first sampled 200 comments—150 of the most egregious "toxic" comments which have a median toxicity rating of "Very toxic" or higher, and 50 "benign" comments that have a median toxicity rating less than "Slightly toxic". We then manually posted these comments to an Instagram account we controlled (with no audience), noting which comments triggered their classifier.

Of the toxic comments, just 41 (27%) triggered the Instagram classifier. These were mostly identity-based attacks (47%), followed by a mix of adult content (15%), profanity (7%), and threats (3%). Two expert raters attempted to label each comment, but we found no unifying themes that might explain why some toxic comments did not trigger detection. Categories reported by our participants for our toxic sample included insults (26%), profanity (22%), and identity attacks (21%). The classifier never triggered on a benign comment. As such, a significant gap remains in the classifier's ability to detect a wide variety of toxic comments.

7 Discussion

Based on our findings, we discuss potential best practices, pitfalls, and paths forward for improving toxic content classifiers to better serve a diversity of perspectives.

Best practices for crowdsourced labeling. During the development of our survey instrument, we were unable to identify any best practices for developing crowdsourcing instruments that gather toxic content ratings. Previous studies used disparate terminology including "abusive", "hateful", "offensive", and "toxic". For sublabeling tasks that involve categorizing toxic content into sexual harassment, identity-based attacks, or insults, we were unable to find terminology or a taxonomy that was evaluated for rater comprehension. Our experiments show that participants solicited from Mechanical Turk in the United States best understood the meaning of "toxic" compared to other terms, and that participants can identify at least five separate categories of toxic content. Given frequent rating disagreement between participants, we also found that five ratings per comment resulted in the best balance between minimizing crowdsourcing costs and achieving a high degree of accuracy. This rating methodology can serve as a future best practice when crowdsourcing labels for toxic content. Furthermore, our results are limited to participants solicited from Mechanical Turk, and should be validated with participants from other crowdsourced platforms.

Towards personalized definitions of toxicity. Our results suggest that personalized tuning of one-size-fits-all models greatly improves the accuracy per user compared to setting a global threshold for all users. In particular, we found that per-user models increased the accuracy of decisions by 86%. These results suggest the feasibility of relying on a general audience for training labels that users then tune to their personal preferences. However, increasing classifier performance beyond this point will remain a challenge without incorporating specific user feedback and examples. An intermediate approach, where models generalize to specific single-trait demographic cohorts rather than individuals, resulted in only a 0–20.6% improvement in accuracy, with age-specific models performing the best. In the absence of personalization or user feedback, platforms might consider increasingly sophisticated, community-based filters that take into account more than just one demographic trait.

Measuring toxicity using existing classifiers. Recent studies in toxic content have begun to leverage toxicity classifiers as a tool for measuring the prevalence of hate and harassment online, with additional post-processing via rater agreement [20, 30, 31]. Given the variations in classifier accuracy across demographic cohorts and types of sites, we caution against off-the-shelf usage of current classifiers without such post-processing or additional calibration. Even at a Perspective toxicity threshold of 0.9 or higher, our participants disagreed with the classifier's verdict in 50% of cases for the sites we measured.

Online Context. Our work does not incorporate the context that a comment is presented in. As such, it may be challenging for a participant to pinpoint if a comment is toxic versus simply sarcastic or joking. We selected this because toxicity detection systems classify text without additional context, and we wanted to evaluate them based on their current usage. Furthermore, users may have different responses to toxic content when they see such context in-situ (e.g., the toxic content may be targeted at an acquaintance). Some areas of future work include understanding how perspectives change if participants are provided with additional context when labeling, identifying if classifiers can be improved by adding context during training, and measuring participant responses to toxic content in-situ of browsing.

8 Conclusion

In this work, we built and deployed a survey instrument to 17,280 participants across the United States and asked them about their perspectives on toxic content online. We found that a participant's attitudes towards filtering toxic content varies across a multitude of factors: their demographic background, their personal experiences with harassment, and even their attitudes towards technology and the state of toxic content online. Given these influences, we showed how personalized tuning of independent thresholds for existing classifiers can improve the accuracy of toxic detection performance by 86% on average, pointing to personalized models as a future area of research in toxic content detection. We have released our labeled toxicity dataset to enable future work in this space and hope that our work presents paths forward for improving toxic content classification for a diverse set of users.

Acknowledgments

The material is based upon work supported by the National Science Foundation under grant #2030859 to the Computing Research Association for the CIFellows Project and through gifts from Google. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of their employers or the sponsors.

References

[1] Amnesty International. Twitter still failing women over online violence and abuse. https://www.amnesty.org/en/latest/news/2020/09/twitter-failing-women-over-online-violence-and-abuse/, 2020.

[2] Anti-Defamation League. Quantifying hate: A year of anti-semitism on twitter. https://www.adl.org/resources/reports/quantifying-hate-a-year-of-anti-semitism-on-twitter#methodology.

[3] R. Arthur. We analyzed more than 1 million comments on 4chan. hate speech there has spiked by 40% since 2015. https://www.vice.com/en_us/article/d3nbzy/we-analyzed-more-than-1-million-comments-on-4chan-hate-speech-there-has-spiked-by-40-since-2015.

[4] J. Blackburn and H. Kwak. Stfu noob!: predicting crowdsourced decisions on toxic behavior in online games. In *Proceedings of the 23rd international conference on World wide web*. ACM, 2014.

[5] L. Blackwell, J. Dimond, S. Schoenebeck, and C. Lampe. Classification and its consequences for online harassment: Design insights from heartmob. In *Proceedings of the ACM on Human-Computer Interaction*, 2017.

[6] L. Blackwell, J. Hardy, T. Ammari, T. Veinot, C. Lampe, and S. Schoenebeck. Lgbt parents and social media: Advocacy, privacy, and disclosure during shifting social movements. In *ACM CHI conference on human factors in computing systems*, 2016.

[7] A. Breland. Why reddit is losing its battle with online hate. https://www.motherjones.com/politics/2019/08/reddit-hate-content-moderation/.

[8] E. Chandrasekharan, M. Samory, A. Srinivasan, and E. Gilbert. The bag of communities: identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017.

[9] D. Chatzakou, N. Kourtellis, J. Blackburn, E. De Cristofaro, G. Stringhini, and A. Vakali. Mean birds: Detecting aggression and bullying on Twitter. In *ACM Web Science Conference*, 2017.

[10] J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. Antisocial behavior in online discussion communities. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[11] D. K. Citron. Addressing cyber harassment: An overview of hate crimes in cyberspace. *Journal of Law, Technology & the Internet*, 2014.

[12] G. Cowan and J. Mettrick. The effects of target variables and settting on perceptions of hate speech. *Journal of Applied Social Psychology*, 2002.

[13] Data & Society. Online harassment, digital abuse, and cyberstalking in america. https://datasociety.net/output/online-harassment-digital-abuse-cyberstalking/, 2016.

[14] T. Davidson, D. Bhattacharya, and I. Weber. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, 2019.

[15] T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. In *AAAI International Conference On Web and Social Media*, 2017.

[16] D. DiFranzo, S. H. Taylor, F. Kazerooni, O. D. Wherry, and N. N. Bazarova. Upstanding by design: Bystander intervention in cyberbullying. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[17] K. Dinakar, R. Reichart, and H. Lieberman. Modeling the detection of textual cyberbullying. In *AAAI International Conference On Web and Social Media*, 2011.

[18] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018.

[19] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. Hate speech detection with comment embeddings. In *The Web Conference*, 2015.

[20] M. ElSherief, S. Nilizadeh, D. Nguyen, G. Vigna, and E. Belding. Peer to peer hate: Hate speech instigators and their targets. In *AAAI International Conference On Web and Social Media*, 2018.

[21] J. Finkelstein, S. Zannettou, B. Bradlyn, and J. Blackburn. A quantitative approach to understanding online antisemitism. In *Proceedings of the AAAI International Conference on Web and Social Media*, 2020.

[22] A. M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. In *12th International AAAI Conference on Web and Social Media*, 2018.

[23] Google Jigsaw. Perspective api. https://www.perspectiveapi.com/#/home.

[24] M. L. Gordon, K. Zhou, K. Patel, T. Hashimoto, and M. S. Bernstein. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *ACM CHI Conferences on Human Factors in Computing Systems*, 2021.

[25] T. Gröndahl, L. Pajola, M. Juuti, M. Conti, and N. Asokan. All you need is "love" evading hate speech detection. In *Proceedings of the ACM Workshop on Artificial Intelligence and Security*, 2018.

[26] A. M. G. Gualdo, S. C. Hunter, K. Durkin, P. Arnaiz, and J. J. Maquilón. The emotional impact of cyberbullying: Differences in perceptions and experiences as a function of role. *Computers & Education*, 2015.

[27] Hatebase. The world's largest structured repository of regionalized, multilingual hate speech. https://hatebase.org/, 2019.

[28] D. Hicks and D. Gasca. A healthier twitter: Progress and more to do. https://blog.twitter.com/en_us/topics/company/2019/health-update.html.

[29] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*, 2017.

[30] Y. Hua, M. Naaman, and T. Ristenpart. Characterizing twitter users who engage in adversarial interactions against political candidates. In *ACM CHI Conference on Human Factors in Computing Systems*, 2020.

[31] Y. Hua, T. Ristenpart, and M. Naaman. Towards measuring adversarial twitter interactions against candidates in the us midterm elections. In *International Conference on Web and Social Media*, 2020.

[32] Instagram. Our progress on leading the fight against online bullying. https://instagram-press.com/blog/2019/12/16/our-progress-on-leading-the-fight-against-online-bullying/.

[33] E. Jain, S. Brown, J. Chen, E. Neaton, M. Baidas, Z. Dong, H. Gu, and N. S. Artan. Adversarial text generation for google's perspective api. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2018.

[34] S. Jhaver, S. Ghoshal, A. Bruckman, and E. Gilbert. Online harassment and content moderation: The case of blocklists. In *Proceedings of the ACM Transactions on Computer-Human Interaction*, 2018.

[35] Jigsaw. Toxic comment classification challenge. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge, 2017.

[36] H. Kwak, J. Blackburn, and S. Han. Exploring cyberbullying and other toxic behavior in team competition online games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015.

[37] J. R. Landis and G. G. Koch. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 1977.

[38] D. Lee. Alphabet-made chrome extension is designed to tune out toxic comments. https://www.theverge.com/2019/3/14/18265851/alphabet-google-jigsaw-tune-chrome-extension.

[39] K. Mahar, D. Karger, and A. X. Zhang. Squadbox: A tool to combat online harassment using friendsourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[40] B. Maher. Can a video game company tame toxic behaviour? https://www.nature.com/news/can-a-video-game-company-tame-toxic-behaviour-1.19647.

[41] S. Melendez. Twitter automatically flags more than half of all tweets that violate its rules. https://www.fastcompany.com/90528941/twitter-automatically-flags-more-than-half-of-all-tweets-that-violate-its-rules?utm_source=morning_brew.

[42] C. Newton. The trauma floor. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona.

[43] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. Abusive language detection in online user content. In *The Web Conference*, 2016.

[44] J. A. Pater, M. K. Kim, E. D. Mynatt, and C. Fiesler. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th International Conference on Supporting Group Work*, 2016.

[45] Pew Research Center. Online harassment 2017. https://www.pewinternet.org/2017/07/11/online-harassment-2017/, 2017.

[46] D. Razo and S. Kübler. Investigating sampling bias in abusive language detection. In *4th Workshop on Online Abuse and Harms*, 2020.

[47] E. M. Redmiles, J. Bodford, and L. Blackwell. "i just want to feel safe": A diary study of safety perceptions on social media. In *International AAAI Conference on Web and Social Media*, 2019.

[48] E. M. Redmiles, S. Kross, and M. L. Mazurek. How well do my results generalize? comparing security and privacy survey results from mturk, web, and telephone samples. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2019.

[49] E. M. Redmiles, Z. Zhu, S. Kross, D. Kuchhal, T. Dumitras, and M. L. Mazurek. Asking for a friend: Evaluating response biases in security user studies. In *25th ACM SIGSAC Conference on Computer and Communications Security*, 2018.

[50] N. Sambasivan, A. Batool, N. Ahmed, T. Matthews, K. Thomas, L. S. Gaytán-Lugo, D. Nemer, E. Bursztein, E. Churchill, and S. Consolvo. "they don't leave us alone anywhere we go": Gender and digital abuse in south asia. In *Proceedings of the Conference on Human Factors in Computing Systems*, 2019.

[51] A. Saravanaraj, J. Sheeba, and S. P. Devaneyan. Automatic detection of Cyberbullying from Twitter. *International Journal of Computer Science and Information Technology & Security (IJCSITS)*, 2016.

[52] K. Thomas, D. Akhawe, M. Bailey, D. Boneh, E. Bursztein, S. Consolvo, N. Dell, Z. Durumeric, P. G. Kelley, D. Kumar, D. McCoy, S. Meiklejohn, T. Ristenpart, and G. Stringhini. Sok: Hate, harassment, and the changing landscape of online abuse. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2021.

[53] US Census. United states census bureau. https://www.census.gov/data.html, 2021.

[54] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, and V. Hoste. Automatic detection and prevention of cyberbullying. In *International Conference on Human and Social Analytics*, 2015.

[55] W. Warner and J. Hirschberg. Detecting hate speech on the World Wide Web. In *Proceedings of the second workshop on language in social media*, 2012.

[56] M. Wich, H. Al Kuwatly, and G. Groh. Investigating annotator bias with a graph-based approach. In *4th Workshop on Online Abuse and Harms*, 2020.

[57] E. Wulczyn, N. Thain, and L. Dixon. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, 2017.

[58] S. Zannettou, T. Caulfield, J. Blackburn, E. De Cristofaro, M. Sirivianos, G. Stringhini, and G. Suarez-Tangil. On the origins of memes by means of fringe web communities. In *Proceedings of the Internet Measurement Conference 2018*. ACM, 2018.

Appendix—Survey instrument

Initial consent form and university contact information
I understand the consent form. I certify that I am 18 years old or older. By clicking the "Yes" button to enter the survey, I indicate my willingness to voluntarily take part in this study.
o Yes
o No

Pre-exercise questions
Today we would like you to help us with this short survey and to review some comments that have been posted online.

What types of sites do you use? [Checkbox]
o Social Networking (Facebook, Twitter)
o Video (YouTube, Twitch)
o News (CNN, Fox, NYT, WSJ)
o Community Forums (Reddit, Craigslist, 4chan)
o Email or messaging (Gmail, WhatsApp, Facebook Chat)

In general, how would you rate the impact of technology on people's lives? [Radio]
o Very positive
o Somewhat positive
o Neutral – neither positive nor negative
o Somewhat negative
o Very negative

Based on your experience, toxic posts or comments are ... [Radio]
o Not a problem
o Rarely a problem
o Occasionally a problem
o Frequently a problem
o Very frequently a problem

Who is the most responsible for addressing toxic posts or comments? [Rank, 1-5]
o Law enforcement
o Users who observe the comment
o The user who receives the comment
o The website hosting the comment
o The user who sent the comment

What does the term "toxic comment" mean to you?
o Open Ended

Labeling exercise
In this section, you will be shown twenty comments taken from various online platforms. Some of these comments may contain toxic content. For each comment, you will be asked a series of questions regarding your perspective on the comment.

[Random comment displayed from Reddit, Twitter, or 4chan. This comment remains visible throughout the subquestions below.]

How toxic do you personally find this comment? [Radio]
o Extremely toxic
o Very toxic
o Moderately toxic
o Slightly toxic
o Not at all toxic

[If any but not at all] Please select all the reasons why below: [Checkmark]
o It is threatening (ex. "If you comment on my feed again, I will personally beat you up")
o It is insulting (ex. "You are an absolutely disgusting hypocrite")
o It is profane (ex. "Why the f*** not you heathen?")
o It is an attack on a person or group based on their identity (ex. "Women are stupid and belong in the kitchen")
o It is sexual harassment (ex. "All I want from you are nudes")
o Other, please specify

Is this a type of comment you would want to see? [Radio]
o I would never want to see it
o It depends on the context

o This is fine for me to see

Should this comment be allowed online? [Radio]
o This comment should be removed
o It depends on the context
o This comment should be allowed

Post-exercise questions
For the sites you use, have you ever seen comments similar to the ones we showed you? [Radio]
o Yes
o No

Have you ever personally been the target of comments similar to the ones you reviewed? [Radio]
o Yes
o No

Is there anything else you would like to tell us about toxic comments?
o [Open ended]

Demographic questions
Which category below includes your age? [Radio]
o Under 18
o 18 - 24
o 25 - 34
o 35 - 44
o 45 - 54
o 55 - 64
o 65 or older
o Prefer not to say

Race [Checkbox]
o White
o Hispanic or Latino
o Black or African American
o Native American or American Indian
o Asian / Pacific Islander
o Other [open ended]
o Prefer not to say

What is your gender? [Radio]
o Female
o Male
o Nonbinary
o Prefer not to say
o Other [Open ended]

Would you describe yourself as transgender? [Radio]
o Yes
o No
o Prefer not to say

What is the highest degree or level of school that you have completed? [Radio]
o Less than high school degree
o High school graduate (high school diploma or equivalent including GED)
o Some college but no degree
o Associate degree in college (2-year)
o Bachelor's degree in college (4-year)
o Master's degree
o Doctoral degree
o Professional degree (JD, MD)
o Prefer not to say
o Other [Open ended]

Do you consider yourself to be: [Radio]
o Heterosexual or straight
o Homosexual
o Bisexual

o Prefer not to say
o Other [Open ended]

How important is religion in your life? [Radio]
o Not important
o Not too important
o Somewhat important
o Very important
o Prefer not to say

Generally speaking, do you consider yourself politically conservative, liberal, independent, or something else? [Radio]
o Conservative
o Liberal
o Independent
o Other [Open ended]
o Prefer not to say

Are you a parent? [Radio]
o Yes
o No
o Prefer not to say

What term have we been using to describe comments in this survey? [Radio]
o Mean
o Abusive
o Bullying
o Toxic

Appendix—Divergent Comments

We show examples of divergence between participant scores and the Perspective API score in Table 8 and Table 9.

Category	Sample	% Comments
Profanity	"Damn cultural colloquialisms being all complicated and stuff"	82%
Negative sentiment	"Song sucks. But I vibe with how he mixed 3 of Lil Wayne videos into one."	6%
Inverted sentiment	"The stupid tongue! :D Cutest thing"	4%
Adult	"penis size shrinking in males exposed to contaminated environments."	2%
Other		6%

Table 8: **High Perspective API Score, Low Groundtruth Score**–The Perspective API over-emphasizes profanity, which participants rarely found toxic.

Category	Sample	% Comments
Identity attack	"We need to actively fight against racism by only targeting white people and letting coloured people get away with crimes"	39%
Political	"And you're getting closer to being locked up behind bars #pediphile #RapistTrump"	17%
Adult	"Hello World! Wanna Hookup; Trade Nude Pics? Visit ...	11%
Threats	"Store them in an unventilated room with hoses that run between the room and your car's exhaust pipe. That'll solve your problem."	9%
Profanity	"the same person that invited you? nofuckingbody	3%
Negative sentiment"	"Sounds like you're a no one who's gonna die bitter and alone and forgotten"	3%
Other		18%

Table 9: **Low Perspective API Score, High Groundtruth Score**–The Perspective API currently misses nuanced identity-based attacks and political attacks that participants deemed toxic.

Appendix—Regression Tables

We present the full parameters and outputs for our quasi-binomial model that predicts the likelihood of rating a random comment as toxic (1 or 0). We show the independent treatment group (either binary or categorical depending on the variable), the reference group, the model coefficient (β), error (*SE*), z-score (z), p-value, and the resultant odds ratio (*OR*). Due to space restraints, we do not present full model results for each individual sublabel model (i.e., whether participant would rate a random comment as an insult, an identity attack, a threat, as profane, or as sexual harassment), and instead direct the reader to the extended version of the paper available at https://arxiv.org/abs/2106.04511.

Demographic	**Treatment**	**Reference**	β	*SE*	z	$Pr(>\|z\|)$	*OR*
Gender	Female	Male	-0.049	0.015	-3.250	0.001	0.952
Gender	Nonbinary	Male	-0.347	0.116	-2.986	0.003	0.707
Age	65 or older	35 - 44	-0.024	0.042	-0.562	0.574	0.977
Age	18 - 24	35 - 44	0.213	0.028	7.488	0.000	1.238
Age	25 - 34	35 - 44	0.204	0.019	10.817	0.000	1.227
Age	55 - 64	35 - 44	-0.020	0.030	-0.665	0.506	0.980
Age	45 - 54	35 - 44	-0.029	0.025	-1.167	0.243	0.972
Race	Minority	Non-minority	0.119	0.016	7.277	0.000	1.126
LGBTQ+	LGBTQ+	Not LGBTQ+	0.497	0.020	25.225	0.000	1.644
Political affiliation	Independent	Liberal	-0.104	0.018	-5.758	0.000	0.901
Political affiliation	Conservative	Liberal	0.024	0.018	1.308	0.191	1.024
Religion	Not too important	Not Important	0.195	0.026	7.617	0.000	1.216
Religion	Somewhat important	Not Important	0.453	0.021	21.947	0.000	1.572
Religion	Very important	Not Important	0.610	0.020	30.177	0.000	1.840
Parent	Yes	No	0.285	0.016	17.360	0.000	1.330
Education	College	High school	0.130	0.026	4.945	0.000	1.139
Education	Advanced degree	High school	0.311	0.030	10.325	0.000	1.365
Impact of Technology	Very negative	Neutral	-0.220	0.080	-2.752	0.006	0.803
Impact of Technology	Somewhat negative	Neutral	-0.140	0.032	-4.357	0.000	0.870
Impact of Technology	Somewhat positive	Neutral	-0.032	0.023	-1.402	0.161	0.968
Impact of Technology	Very positive	Neutral	0.133	0.025	5.318	0.000	1.142
Toxic Content a Problem?	Rarely a problem	Not a problem	0.029	0.034	0.863	0.388	1.030
Toxic Content a Problem?	Occasionally a problem	Not a problem	-0.043	0.032	-1.314	0.189	0.958
Toxic Content a Problem?	Frequently a problem	Not a problem	0.028	0.033	0.848	0.397	1.029
Toxic Content a Problem?	Very frequently a problem	Not a problem	0.117	0.037	3.188	0.001	1.125
Party most responsible	Law Enforcement	Bystander	0.248	0.035	7.093	0.000	1.282
Party most responsible	User who Receives	Bystander	-0.334	0.032	-10.427	0.000	0.716
Party most responsible	Hosting Platform	Bystander	-0.348	0.028	-12.481	0.000	0.706
Party most responsible	User who sent the comment	Bystander	-0.480	0.027	-17.973	0.000	0.619
Witnessed Toxic Content	True	False	-0.249	0.018	-14.208	0.000	0.779
Experienced Toxic Content	True	False	0.394	0.017	23.547	0.000	1.482

Table 10: **Toxicity Model**—Logistic regression showing the likelihood a participant will flag a random comment as toxic.

Why They Ignore English Emails: The Challenges of Non-Native Speakers in Identifying Phishing Emails

Ayako A. Hasegawa
NTT

Naomi Yamashita
NTT

Mitsuaki Akiyama
NTT

Tatsuya Mori
Waseda University / NICT / RIKEN AIP

Abstract

Prior work in cybersecurity and risk management has shown that non-native speakers of the language used in phishing emails are more susceptible to such attacks. Despite much research on behaviors English speakers use to avoid phishing attacks, little is known about behaviors of non-native speakers. Therefore, we conducted an online survey with 862 non-native English speakers (284 Germans, 276 South Koreans, and 302 Japanese). Our findings show that participants, especially those who lacked confidence in English, had a higher tendency to ignore English emails without careful inspection than emails in their native languages. Furthermore, both the German and South Korean participants generally followed the instructions in the email in their native languages without careful inspection. Finally, our qualitative analysis revealed five main factors that formed the participants' concerns in identifying English phishing emails. These findings highlight the importance of providing non-native speakers with specific anti-phishing interventions that differ from those for native speakers.

1 Introduction

Phishing is a form of online fraud that acquires such sensitive information as account credentials and credit card information by masquerading as a legitimate business or reputable person. Since the mid-90s, an increasing body of research in the fields of cybersecurity and risk management has led to the development of techniques to combat phishing [15, 56]. However, it remains a huge cybersecurity threat [64]. It is noteworthy that COVID-19 has caused a further massive increase in phishing attacks [36].

Among Internet users, non-native speakers of the language used in phishing emails are more susceptible to such attacks. Recent work has shown that English proficiency level significantly affects the ability of the users to identify English phishing emails, and evidently lower English proficiency levels lead to increased phishing susceptibility [2]. Although research shows that non-native speakers are more susceptible to phishing attacks, little is known about their coping behaviors: we still lack an understanding of the differences in behavior (or reactions) of people when they receive an email containing instructions to click on a URL link or open an attachment in a non-native language compared with when they receive it in their native languages.

In the fields of psychology and cognitive science, research has shown that people tend to behave differently (i.e., perform more poorly) when using a non-native language compared with their native language [51, 61]. However, research suggests somewhat incongruous results. First, previous research on risk and uncertainty shows that people tend to be risk-averse in the face of uncertainty [54]. This implies that people may become more risk-averse when dealing with emails written in a non-native language because they are less confident about being able to identify phishing attacks written in a non-native language. As a result, people may simply ignore such emails without careful inspection. In contrast, other research has shown that people tend to make more risk-prone decisions when using a non-native language [13, 34]. Therefore, a non-native speaker may follow the instructions written in the email without careful inspection, which could lead to unwanted consequences, such as breaches of critical sensitive information.

To help non-native speakers defend themselves against phishing attacks, we must understand their current practices of dealing with emails written in a non-native language. In particular, we are interested in understanding the concerns of non-native speakers when they are involved in phishing attacks and the differences in the ways they deal with emails

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

(with links and attachments) depending on whether they are written in their native language or in a non-native language.

In order to explore behavioral tendencies of non-native English speakers (NNESs) and their concerns about identifying English phishing emails, we conducted an online study with 862 NNESs (284 Germans, 276 South Koreans, and 302 Japanese) who are full-time workers exposed to the risks of phishing attacks in English. We recruited NNESs because English remains the dominant language on the Internet [60]. Although German, South Korean, and Japanese peoples are all NNESs, their average English proficiency levels differ; Germans have the highest, while the Japanese have the lowest [20]. We studied participants' behavior toward emails and their confidence and concerns about identifying English phishing emails.

Our findings indicate that participants adopted more security-risk-averse behaviors (i.e., ignoring emails without careful inspection) when the emails were written in English rather than in their native languages. This tendency was salient for those who lacked confidence in reading English. Furthermore, both the German and South Korean participants generally adopted more security-risk-prone behaviors (i.e., following the instructions in the email without careful inspection) when the emails were written in their native languages than the Japanese participants. In addition, qualitative analysis of their open-ended answers revealed five main factors that formed their concerns in identifying English phishing emails, which differ from the concerns they have in identifying phishing emails in their native languages. These findings highlight the importance of providing non-native speakers with specific anti-phishing interventions that differ from those for native speakers.

This study makes the following contributions:

1. This work is among the first that systematically explores the relationship between users' English proficiency levels and their reactions/behaviors when receiving an email that includes links and attachments.
2. Our results show that users have specific concerns about identifying a phishing email written in their non-native language (English) and that they adopt different strategies when receiving emails written in their native and non-native languages.
3. Our findings provide design implications that help users combat phishing attacks in their non-native languages (English).

2 Background and Research Questions

In this section, we review the literature that is closely related to this study. We first review studies that explored the factors that influence users' susceptibility to phishing emails. Next, we review previous works that examined the effect of users' language and culture on their susceptibility to phishing emails. Finally, we highlight the research questions of this study.

2.1 Factors Related to Phishing Susceptibility

The factors related to susceptibility to phishing (including spear phishing) emails found by previous studies can be classified into three categories: (i) user demographics, (ii) anti-phishing strategies, and (iii) contents and contexts of phishing.

User Demographics. Many researchers have found that basic demographics such as age and gender are related to phishing susceptibility [32, 38, 47, 57]. However, some studies yielded incongruous results because they used different methods and studied different populations. For example, Sheng et al. [57] reported young people were most susceptible to phishing whereas Li et al. [38] concluded that older people were the most susceptible. Research has also revealed that users' attributions or traits such as personality traits (Big Five) [3, 25], cognitive impulsivity [5, 49], employment department and position [38], and education level [42, 49] were related to phishing susceptibility. In terms of user skills, studies have reported that user security knowledge, awareness, behavior, and previous anti-phishing training experience were significantly related to phishing susceptibility [6, 22, 27, 57]. Vishwanath et al. [65] found that a heavy email load (i.e., the number of received emails) had a strong and significant influence on phishing susceptibility.

Anti-phishing Strategies. Several studies indicated that people often did not pay attention to reliable phishing cues and their strategies failed to identify phishing emails or suspicious URLs [1, 3, 18, 27, 48, 53]. For instance, Downs et al. [18] reported that participants in their study used various strategies to determine the validity of emails, primarily centered around interpreting the email text rather than focusing on more reliable phishing cues in headers or the URLs associated with the links. On the other hand, Vishwanath et al. [65] and Wang et al. [67] found that individual attention to email sources, grammatical errors, and misspellings were significantly negatively related to phishing susceptibility. They also concluded that individual attention to urgent cues and subject lines were significantly positively related to phishing susceptibility.

Contents and Contexts of Phishing Emails. Given that users' anti-phishing strategies center around interpreting email texts, researchers have studied how users' behaviors are affected by email contents. Researchers have classified the contents of phishing emails based on a seminal work by Cialdini [12], which identified principles that triggered people's decisions to comply with requests (called "principles of persuasion"). They found that the presence of authority cues [5, 68], consistency [63], and scarcity [63] increased users' phishing susceptibility. From the viewpoints of contexts, participants are more susceptible when phishing messages are specific to their situations [24, 28, 30]. Unsurprisingly, several studies revealed that the contents and context of phishing emails to which participants were more susceptible depend on demographics of participants [38, 39, 47].

2.2 Impact of Culture and Language

Culture. Cross-cultural studies are positioned as a crucial theme in the field of cybersecurity because culture directly impacts security-related phenomena [14]. Recently, many researchers have conducted a variety of cross-cultural security studies, such as those on the security behavior intentions scale (SeBIS) [55], generated passwords [43], smartphone unlocking [26], and account security incident response [52]. Some of these cross-cultural studies adopted Hofstede's cultural dimensions [29] to interpret the observed differences in security behavior by linking them to the national characteristics, such as the individualism-collectivism dimension [52].

The cross-cultural approach has also attracted interest in phishing research. Butavicius et al. [6] and Tembe et al. [62] recruited participants from multiple countries and showed that those with higher individualism scores (e.g., the U.S. participants) were less likely to be phished. Both works suggest that low levels of individualism may fuel a desire to respond to requests from others to maintain group harmony, which includes requests in phishing emails (especially from an authority figure). Flores et al. [22] reported that factors (individual demographics) that were significantly correlated with phishing susceptibility differed among countries.

Language. Although language is known to have a considerable influence on one's thoughts and behavior, few studies in cybersecurity have focused on language. A broad body of research in the fields of psychology and cognitive science shows that people face various interpretation and reasoning problems when using a non-native language [11, 51, 61, 66]. For instance, Takano and Noda [61] demonstrated that using a foreign language caused a temporary decline in thinking task performance. Rear [51] compared the critical thinking skills of Asian students in their native language and English contexts and argued that using a foreign language considerably interfered with critical thinking. Some researchers have also identified problems that non-native speakers face during Internet use, such as online searches [11]. On the other hand, some psychological researchers demonstrated that using a foreign language reduced decision-making bias, that is, the loss aversion bias that people have in their native language contexts was reduced in foreign language contexts [13, 34].

Although studies have explored the impact of culture on users' phishing susceptibility, the impact of language (especially language barriers) on this phenomenon is not yet fully understood. So far, little work has addressed the impact of language barriers of NNESs on their susceptibility to phishing emails. Among the few studies that investigated language issues in cybersecurity, Alseadoon et al. [2] and Kävrestad et al. [33] conducted a phishing identification task in Saudi Arabia and Sweden, respectively. They revealed that the NNESs' self-perceived English proficiency level significantly affected their ability to identify phishing English emails [2] and legitimate English emails [33], respectively.

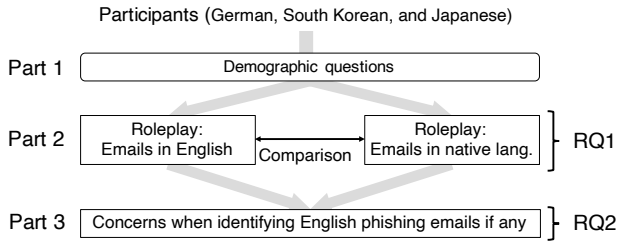


Figure 1: Overview of our survey design and research questions (RQs).

2.3 Research Questions

In summary, although previous works suggest that NNESs may be more susceptible to phishing attacks, it remains unclear how language affects non-native speakers' strategies to combat phishing attacks. To help non-native speakers defend themselves against phishing attacks, it is critical to understand their current practices of dealing with emails written in their non-native language. In this paper, we ask:

RQ1. Do NNESs show different behavioral tendencies (e.g., security-risk-prone vs. security-risk-averse) toward native language and English emails?

RQ2. What are the NNESs' concerns about identifying English phishing emails?

Following the suggestion of Lastdrager et al. [37], who addressed anti-phishing interventions designed specifically for children, we advocate for anti-phishing interventions designed specifically for NNESs.

3 Methods

We designed an online survey to understand NNESs' behavioral tendencies (**RQ1**) and concerns (**RQ2**) about English phishing emails.

3.1 Survey Design

Figure 1 summarizes the design of our survey and corresponding research questions[1]. Our survey consisted of three parts. In Part 1, we asked the participants about their demographics; then in Part 2, we explored their behavior and attention toward emails (**RQ1**); and finally in Part 3, we asked the participants with low confidence in identifying English phishing emails about their concerns. (**RQ2**).

In Part 2, randomly selected half of the participants from each country were shown a set of English emails that included phishing emails. The other half were shown the same set of emails translated into their native languages. All participants were provided with a scenario that described the background

[1] The entire study was approved by our Institutional Review Board.

of receiving the emails and asked how they would respond to them. To minimize the effects of email content, we adopted a between-subjects design for Part 2. In Part 3, all participants were asked about their past experiences of being deceived by phishing emails and their confidence in identifying phishing emails in their native languages and in English. We adopted a within-subject design for Part 3 because we were interested in understanding whether people had different experiences and confidence levels when the language of the emails differed. Furthermore, it is worth noting that we conducted Part 3 after Part 2 because we were concerned that the participants' behavior (in Part 2) may be affected if they knew the focus of our study was phishing (as revealed in Part 3). A previous study showed that revealing such information would improve the participants' performance to identify phishing emails during the experiment [49]. We did not inform the participants that they were participating in a phishing study in Part 2. The demographics questions in Part 1 also did not include questions about their phishing experiences.

The questionnaire items (including the email materials) were translated into German, Korean, and Japanese by two professional translators of the respective language to ensure their validity.

3.2 Procedure

In this subsection, we introduce the procedure of our study and the preventive measures that protected the privacy of our participants during the study.

Screening Survey. To recruit eligible participants, we implemented a short screening survey prior to our main survey. The screening survey included four demographic questions: age, self-identified gender, occupational status, and native language. In the middle of the screening survey, we asked an attention check question. Those who were deemed eligible for our survey (participation eligibility is described in Section 3.4) proceeded to our survey. Our screening survey included a consent form and instructions. In the instructions, participants were provided the survey title, estimated time, compensation, and confidentiality of the survey data. Our survey title was "Survey of emails written in <participants' native language> or English". Based on other security-related studies that conducted online surveys [1, 46], we did not use security-related terms (e.g., phishing) in either the survey title or instructions to avoid recruiting biased participants who were only interested in computer security.

All participants were required to complete consent forms before starting the main survey (Parts 1 to 3).

Part 1: Demographics. In addition to the basic demographic questions from the screening survey, we asked the following six questions: education level, whether they were IT professionals, confidence level in their English reading skills (6-point Likert scale), total years spent learning English, and the average number of emails they received each working day in their native language and English.

Part 2: Behavioral Tendencies (RQ1). Based on previous phishing studies, we measured the participants' behaviors based on their performances in a scenario-based roleplay task [7, 8, 18, 49, 50, 57]. The roleplay enables researchers to study phishing without conducting an actual simulated phishing attack [57].

We first gave participants fictitious profile information about the email recipient for roleplay in their native language. We then showed screenshots of four emails that included phishing emails and asked them to answer how they would respond if they received each email by selecting provided options. At the end of Part 2, we asked the participants about the email elements to which they usually paid attention when they received emails. They chose their top 3 email elements from a list of representative elements (e.g., sender's email address, subject line, grammatical errors and misspellings), which were adopted from Vishwanath et al. [65].

Part 3: Confidence and Concerns about English Phishing Emails (RQ2). In Part 3, we asked participants questions about their confidence and concerns about identifying English phishing emails. To avoid misunderstandings, we defined "phishing attacks" at the beginning of Part 3. Then, we asked how often they received both work-related and personal suspicious emails (except company phishing training). We specifically asked them how often they received such emails written in both their native language and English.

Next, we asked about their experiences of being deceived by phishing in both work-related and personal emails, except for training. We asked about clicking on a link or opening an attached file in phishing emails written both in their native language and English, regardless of the damage.

Participants were then asked about their confidence levels for identifying phishing emails. They assessed their agreement or disagreement with these two statements: "I can always identify a phishing email written in <participant's native languages>" and "I can always identify a phishing email written in English" on a 6-point Likert scale from "strongly disagree" to "strongly agree." Depending on their answers about their level of confidence, participants were asked either why they thought that they could or could not identify English phishing emails in an open-ended question.

For a manipulation check, we included a question in the middle of Part 3 that asked about the definition of phishing. This was to confirm their understanding of phishing emails. They were asked to choose the best definition of phishing from three options: the definitions of phishing, ransomware, and distributed denial-of-service (DDoS). An attention check question was also included in the middle of Part 3. Participants who answered either the definition check or the attention check incorrectly, or both, were excluded from our dataset to ensure the quality of our analysis results. All participants received compensation, even if they did not pass these checks.

3.3 Materials for Roleplay Task

For our roleplay task, we carefully examined previous phishing studies as mentioned in Section 3.2 and finally prepared four emails: an *obvious-phishing* email, two *uncertain* emails, and a *genuine* email (Table 1). As shown in Fig. 2, all used screenshots of the emails follow the format of Gmail. The obvious-phishing email contained features that appeared to be undeniably illegitimate. The genuine email contained no features that suggested phishing. The uncertain emails contained some features that suggested the possibility of phishing; however, such information alone did not provide sufficient evidence to identify whether the email was phishing based only on the appearance of the screenshots. The contents of the emails of our roleplay task must resemble those received by NNESs on a daily basis. If NNESs receive an email in English from a service that is unavailable in their country, they are likely to ignore it based on the unnatural context. Therefore, the senders (or spoofed senders) of the emails must be well-known, worldwide services (e.g., PayPal and LinkedIn) or business acquaintances to increase the feeling of verisimilitude in NNESs about an email in English. The obvious-phishing and uncertain emails were collected from an online archive of phishing emails (MillerSmiles.co.uk [41]) and the dataset used by Canfield et al. [7, 8]. The genuine email was taken from an inbox of one of the authors. We then arranged them for this study (e.g., displayed names and dates). We kept the survey short by providing a limited number of emails for this roleplay task that could be completed in a few minutes in order to reduce participants' fatigue.

Recent studies examined spear phishing emails applying the psychological principles of persuasion [47, 63] as mentioned in Section 2.1. Instead of covering various scenarios concerning such psychologically persuasive contexts, our roleplay task focuses on phishing cues that might be fundamental metrics when users identify phishing emails. As summarized in Table 1, the obvious-phishing and uncertain emails contained two or more features often associated with such practices as phishing cues: suspicious sender email addresses, suspicious URLs, impersonal greetings, hidden URLs, attached files, requests for sensitive information, and requests requiring urgent action [7, 8, 18, 59]. We did not use an email that contained obvious grammatical errors or misspellings. This is because it would be impossible to replicate them accurately across languages and we wanted to minimize experimental variability. In the obvious-phishing email (b) (Table 1), a suspicious URL was displayed, which Canfield et al. [8] described as the most valid cue for identifying phishing emails. The sender's email address in phishing email (b) was also suspicious. Although the name of a well-known service (LinkedIn) appeared in the URL and in the sender's email address, their positions in the URL and email address structure were inauthentic. The uncertain email (a) contained a hidden URL, and the URL was hidden by an HTML button that displayed text. The URL must be uncovered by hovering over it with a mouse to identify whether it was phishing or genuine. In the uncertain email (d), a zip file was attached, which could contain harmful files such as malware.

Table 1: Features of four emails used in roleplay task.

	Sender	Legitimacy	Phishing cues
(a)	PayPal (Card expiration notice)	Uncertain	· Impersonal greeting · Hidden URL (HTML button) · Request for sensitive information · Request for urgent action
(b)	LinkedIn (Login notification)	Obvious phishing	· Suspicious sender's email address · Impersonal greeting · Suspicious URL · Request for urgent action
(c)	Coworker (Meeting invitation)	Genuine	N/A
(d)	IT service staff (Alert notice)	Uncertain	· Attached zip file · Request for urgent action

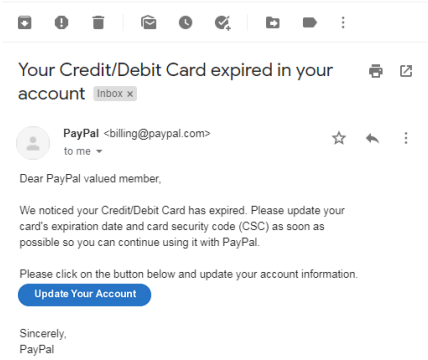


Figure 2: Example of email screenshots shown to participants in the roleplay task (Email (a), English version).

The participants played the role of a male employee who was given a prevalent name in each country (e.g., Japanese participants were given an identity of "Taro Yamada") working at the ABC Company. Each participant was informed that he had PayPal and LinkedIn accounts, which, respectively, corresponded to the senders of emails (a) and (b). We also informed the participants of the email addresses of his boss and the company's IT staff, which, respectively, corresponded to the senders of emails (c) and (d).

We then provided the following four options to participants in each email: "I'd ignore it without referring to any other information than this screenshot;" "I'd follow its instruction without referring to any other information than this screenshot;" "I'd refer to some other information than this screenshot to decide how to respond[2];" and "Other." Although, in reality, users could perform multiple actions (e.g., they ignore an email after checking the validity of sender address), to reduce the complexity of our user study, our roleplay tasks ask participants to choose their *initial reaction* rather than an email response procedure. We also asked the participants who

[2]Hovering over links is included in this option.

chose"I'd refer to some other information..." or "Other" to specify the information they would refer to or what actions they would take in an open-ended form. Since we wanted to protect our participants from accessing phishing websites but did not want them to think that those emails are phishing, we asked them to answer the questions without searching any information contained in the emails.

Although we designed our roleplay tasks according to the aforementioned prominent literature, the ecological validity of the anti-phishing study of user behavior needs further improvement (Please see Section 5.2).

We provide the full questionnaire in Appendix A.

3.4 Participants

We recruited participants from three non-native English-speaking countries. The English skills of the citizens of these countries and their confidence in English might affect their responses to English emails. We used the EF English Proficiency Index (EF EPI) 2019 [20] and selected one country from each English proficiency level group: Germany from with the very high or high level, South Korea from the moderate level, and Japan from low or very low level.

In each country, we limited the participants to full-time workers who were at least 18 years old and native speakers of the country's official language (e.g., German samples only consisted of native German speakers). We recruited workers to improve the ecological validity of our study. For NNESs, workers face higher potential risks of phishing attacks written in English because they are more likely to be exposed to English than non-workers. We recruited a broad array of participants with quota sampling to match the demographics of working populations.

We recruited participants and conducted our survey through a survey company (Macromill [40]) that has large-scale, global online panels. The participants received a compensation, which roughly equals US$4.7. This survey was done in July and August, 2020.

We analyzed valid responses from 862 participants: 284 Germans, 276 South Koreans, and 302 Japanese. Participants finished our survey in 7.5 minutes (median), including the screening survey. Table 2 shows the demographics of our participants. Their age and gender distributions were similar in all three countries. Although there were some differences among the three countries in demographics other than age and gender, the distributions of the demographics between the two groups divided by language in our roleplay task (native language group and English group) were similar in each country. The percentages of participants who were confident in their English reading skills were high in Germany, followed in descending order by South Korea and Japan.

Table 3 shows the frequency that participants received suspicious emails and their experience being deceived by phishing emails. Although the frequency of receiving suspicious emails was lower in English than in their native languages in all three countries, at least a quarter of the participants received suspicious emails in English at least once a month. This indicates that NNESs are regularly exposed or perceive to be exposed to English phishing emails. Although the percentages of participants who have been deceived by phishing emails in English was also lower than in their native language, this result may be influenced by the fact that they obviously receive more suspicious emails in their native languages than in English. In this paper, we explore the differences in users' susceptibility between the contexts of their native languages and English when they actually receive a phishing email.

3.5 Data Analysis

In this study, we conducted two types of data analysis: participants' behavioral tendencies toward emails in our roleplay task (**RQ1**) and their concerns about identifying phishing emails in English (**RQ2**).

First, we categorized participants' behaviors toward phishing in a roleplay task based on two typical indexes introduced in Section 1: security-risk-prone and security-risk-averse behavioral tendencies.

- Security-Risk-Prone Behavior. We defined security-risk-prone behavior as the participants following the instructions from the sender (e.g., clicking a link or opening an attached file) without any inspection. This behavior is problematic when the email is likely phishing. In our analysis, we counted the participants who answered, "I'd follow its instruction without referring to any other information than this screenshot," to the obvious-phishing or uncertain emails ((a), (b), and (d) in Table 1).
- Security-Risk-Averse Behavior. We defined security-risk-averse behavior as the participants ignoring instructions from the sender without any inspection, even when the email is likely genuine. We counted the participants who answered, "I'd ignore it without referring to any other information than this screenshot," to the genuine or uncertain emails ((a), (c), and (d) in Table 1).

For each email, we tested whether there was a significant difference between the percentages of participants who engaged in risk-prone behavior toward emails in their native language and those in English (Chi-square tests with Bonferroni correction for multiple comparisons). We tested security-risk-averse behavior in the same manner. Furthermore, to explore factors that affect an individual's security-risk-prone/risk-averse behavioral tendency, we performed ordinal logistic regression analyses. Specifically, we used the following model: security-risk-prone/averse behavior $\sim$ age group + IT expertise + confidence in reading English + Email load + culture. These independent variables were selected based on the findings of the existing literature [38, 57, 65]. We confirmed that each pair of our independent variables had no multicollinearity and that the proportional odds assumption was satisfied.

Table 2: Basic and extensive demographics of our participants.

Country		Germany		South Korea		Japan	
Language used in our roleplay task		German (*N*=140)	English (*N*=144)	Korean (*N*=141)	English (*N*=135)	Japan (*N*=148)	English (*N*=154)
Age	18-29	24.3%	20.8%	29.8%	25.9%	23.6%	25.3%
	30-39	24.3%	27.1%	22.0%	25.9%	25.7%	21.4%
	40-49	26.4%	27.1%	24.1%	24.4%	25.0%	26.6%
	50-59	22.1%	20.8%	19.1%	19.3%	23.0%	20.1%
	60 or over	2.9%	4.2%	5.0%	4.4%	2.7%	6.5%
Self-identified gender	Male	55.0%	56.3%	59.6%	57.8%	60.1%	55.8%
	Female	45.0%	43.8%	40.4%	42.2%	39.9%	44.2%
Level of education	No high school/High school	25.0%	23.6%	9.9%	11.1%	19.6%	26.0%
	Assoc. degree/Tech. degree	39.3%	45.1%	12.8%	9.6%	20.3%	18.8%
	Bachelor's degree	33.6%	29.2%	71.6%	68.9%	52.7%	46.1%
	Graduate degree	2.1%	2.1%	5.7%	9.6%	7.4%	8.4%
IT professionals	% Professionals	30.7%	27.1%	19.9%	19.3%	8.1%	6.5%
Confidence of English-reading	% Positive (6-point scale)	77.9%	77.8%	41.1%	42.2%	12.2%	14.3%
Years of English learning	Ave.	8.9	8.4	11.6	12.4	8.4	8.2
Received emails per day	Ave.: Native language	25.3	23.1	13.5	14.2	28.7	33.1
Received emails per day	Ave.: English	6.5	4.7	3.0	2.7	1.4	2.2

Table 3: Participants' experiences with phishing emails.

Country	% Participants who receive suspicious emails at least once a month		% Participants who have been deceived by phishing emails	
	Native	English	Native	English
Germany	58.1%	47.5%	25.4%	14.1%
South Korea	56.7%	39.1%	14.5%	10.1%
Japan	51.7%	26.5%	6.0%	1.3%

Next, we explored the participants' concerns about identifying English phishing emails in open-ended questions. Original open-ended comments were collected in participants' native languages and professional translators translated them into English. Two independent coders then rated them through an inductive thematic analysis method, which identifies, analyzes, and reports patterns (themes) within data [4]. The coders practiced rating a subsample of users' responses and discussed differences until they reached a consensus before rating the remainder of the data. Because participants sometimes provided multiple concerns, we allowed multiple themes per response. Accordingly, we calculated the inter-rater reliability using the Kupper-Hafner statistic [21].

4 Results

In this section, we aim to answer our research questions by analysing the results of our roleplay task and survey questions. We first addressed **RQ1** by studying participants' behavior toward emails. Next, we addressed **RQ2** by studying participants' confidence and concerns about identifying English phishing emails. This study aims to unveil the differences in participants' behavior and perceptions between their native languages and English. Please note that national or cultural differences are out of our scope (see 5.2 for more details).

4.1 RQ1: Behavior toward the Emails

4.1.1 Security-Risk-Prone/Averse Behavioral Tendency

Table 4 shows the percentages of the participants with security-risk-prone behavioral tendencies for each email in our roleplay task. In Germany and South Korea, the percentage of participants who engaged in security-risk-prone behaviors was lower in English contexts than in their native language contexts. Especially in South Korea, the difference was large and statistically significant for all three emails (p<.05). In contrast, in Japan, more participants engaged in security-risk-prone behavior in English contexts than in Japanese contexts. In all three countries, the percentages of participants who engaged in security-risk-prone behavior in response to the obvious-phishing email (b) with the suspicious URL and to the uncertain email (a) were similar. Participants did not seem to look for and rely on a suspicious URL for their decision-making, although Canfield et al. [8] described it as the most valid cue for identifying phishing emails.

As Table 4 shows, in all three countries, the percentage of participants who engaged in security-risk-averse behavior was higher for emails in English than for emails in their native languages. For the genuine email (c), which contained no phishing cues, 24–29% of the participants engaged in security-risk-averse behavior in English contexts. We found that the differences in the percentages of participants with security-risk-averse behavior between their native language and English contexts were larger for email (a), which was sent from PayPal, than emails (c) and (d), which were respectively sent from the coworker and company staff. This tendency was common in all three countries. In other words, participants appear more likely to ignore an email from a service with which they have no personal relationship than one from a sender with whom they have an established relationship in English contexts. This result suggests that the expected ex-

Table 4: Percentage of participants who engaged in security-risk-prone/averse behaviors in our roleplay task.

Behavioral tendency	Country	Language	(a) Uncertain	(b) Obvious phishing	(c) Genuine	(d) Uncertain
Security-risk-prone behavior	Germany	Native	27.1%	30.7%	N/A	53.6%
		English	22.2%	23.6%		45.8%
	South Korea	Native	**41.8%***	**43.3%***		**50.4%***
		English	**25.9%***	**25.2%***		**31.1%***
	Japan	Native	8.8%	8.8%		23.6%
		English	14.3%	14.9%		24.7%
Security-risk-averse behavior	Germany	Native	48.6%	N/A	19.3%	32.1%
		English	61.8%		24.3%	40.3%
	South Korea	Native	**20.6%****		21.3%	30.5%
		English	**41.5%****		28.9%	37.0%
	Japan	Native	44.6%		18.9%	46.6%
		English	56.5%		29.2%	48.7%

Bold font indicates that the difference between the two groups (native language and English) is statistically significant (Chi-square tests). Significance levels are ***p<.001; **p<.01; *p<.05, whose p-values are corrected for multiple testing using the Bonferroni method.

Table 5: Regression analysis for Security-risk-prone/averse behavioral tendencies in English contexts.

Independent variables	Security-risk-prone				Security-risk-averse			
	Coefficients	Std. Err.	p-values		Coefficients	Std. Err.	p-values	
Age group	-.2702	.0814	<.001	***	.2068	.0745	.0055	**
IT professional	.2602	.2555	.3086		-.1440	.2375	.5443	
Confidence in reading English	.2504	.0867	.0039	**	-.3263	.0844	<.001	***
Num. Received English emails	-.0234	.0152	.1245		-.0294	.0159	.0641	
Korean	-.0717	.2362	.7616		-.3981	.1176	.0473	*
Japanese	-.5300	.2794	.0578		.0635	.2068	.7589	

p-values test the hypothesis that coefficients are zero, i.e., independent variables do not affect security-risk-prone/averse behavior. Significance levels are ***p <.001; **p <.01; *p <.05. Security-risk-prone/averse behavior (dependent variables): the number of emails the participants engaged in security-risk-prone/averse behaviors (0 to 3). Age group: 18-29, 30-39, 40-49, 50-59, or ≥60. IT professional: professional or non-professional (we set the non-professional as the baseline). Confidence in reading English: 6-point Likert scale 0 (strongly disagree) to 5 (strongly agree). Culture: Germany, South Korea, or Japan (we set Germany as the baseline).

tent to which relationships are impacted by ignoring emails may be negatively related to security-risk-averse behavior in English contexts.

As shown in Table 5, age groups and confidence in reading English had significant effects on participants' security-risk-prone/averse behaviors in English contexts; younger participants with more confidence in reading English were more likely to follow the instructions in obvious-phishing and uncertain emails written in English, and they were also less likely to ignore the instructions in genuine and uncertain emails written in English. The result indicating that factors related to participants' self-perceived English proficiency level significantly affect their behavior toward English emails is consistent with previous phishing studies that examined NNESs [2, 33]. Contrary to the expectations from previous phishing studies [22, 57, 65], IT expertise and the number of received emails did not significantly affect participants' behavior in English contexts. A prior work [45] in language communication reported that NNESs' behavior was more influenced by their self-perceived English fluency than objective English fluency. Our results seems to support that conclusion: confidence in English reading (self-perceived English index) did affect NNESs' behavior more than the number of received English emails (the objective English index relates to familiarity with English).

4.1.2 Participants' Inspection Behaviors

We analyzed the open-ended responses of participants who reported that they would refer to some other information than the screenshot. The most frequent inspection behaviors were the same regardless of whether the participants were shown emails in English or their native language: participants would log in to the website without clicking the link in the email, which is generally recommended as an anti-phishing reaction [44], for emails (a) and (b), and they would ask their coworkers for emails (c) and (d). In all three countries, the percentages of participants who reported that they would use Internet search were higher in their native language environment than in English. In English contexts, 0.4% of German, 4.8% of Korean, and 6.5% of Japanese reported that they would use an online translator (on average of four emails). Please see Appendix B for more details.

4.1.3 Attention to Email Elements

Previous studies found that individual attention to email sources and grammatical errors/misspellings were significantly and negatively related to phishing susceptibility, and that attention to urgency cues and subject lines were significantly and positively related to phishing susceptibility [65,67]. Fig. 3 shows the percentages of the participants who usually

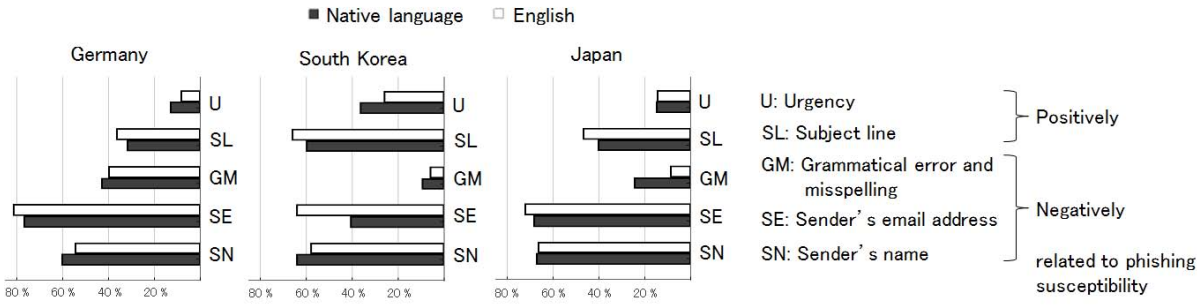


Figure 3: Email elements to which our participants pay attention.

paid attention to each email element in their native language and English contexts. Although there were some differences in elements that participants paid attention to between the three countries, we focus on the common differences between their native languages and English. In all three countries, participants paid less attention to grammatical/misspelling errors and more attention to the sender's email addresses and subject lines in English contexts. This indicates that participants tended to rely more on information recognized at a glance to roughly grasp the content and context of the emails in English contexts. We note that excessive reliance on subject lines for phishing identification is risky because they often serve as a lure in phishing emails [65]. In countries with relatively low English proficiency (i.e., Korea and especially Japan), the percentage of participants who focused on grammatical errors and misspellings was markedly lower in English contexts than in native language contexts. This reflects the fact that participants with low confidence in their English reading skills believed that they were unable to detect such errors in English.

4.2 RQ2: Confidence and Concerns about English Phishing Emails

4.2.1 Confidence in Identifying Phishing Emails

Figure 4 shows participants' degree of confidence in identifying phishing emails. The percentages of participants who were not confident in identifying English phishing emails (i.e., answered "strongly disagree" to "somewhat disagree.") were 24.6% (70/284) in Germany, 43.5% (120/276) in Korea, and 60.0% (181/302) in Japan. In all three countries, the percentage of participants who were confident that they can identify phishing was lower in English than in their native language. We conducted a correlation analysis and found that participants' degree of confidence in identifying English phishing emails was positively correlated with their confidence in their English reading skills (Germany ρ=.378, p<.001; South Korea ρ=.456, p<.001; Japan ρ=.452, p<.001). Conversely, following the several previous studies that showed that the

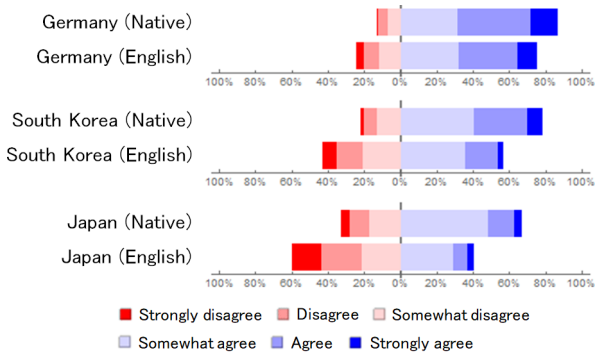


Figure 4: Participants' confidence in identifying phishing.

participants' confidence in identifying phishing did not explain their phishing identification performance in the native language contexts [17, 31, 46], we also found that confidence in identifying English phishing was not significantly correlated with participants' security-risk-prone/averse behavioral tendencies in English contexts in any of the three countries.

4.2.2 Concerns about Identifying Phishing in English

Of the 862 participants, 371 participants who were not confident in identifying English phishing emails (as described in Section 4.2.1) were asked about their concerns. Since we aimed to explore their specific problems to determine anti-phishing interventions for NNESs, we excluded 144 unclear responses such as "Because I am not good at English." As a result, we conducted thematic analysis of 227 responses and found five main concerns: *difficulty understanding English email content* (70.0%), *difficulty identifying errors and unnatural language in English* (15.0%), *unfamiliarity with English phishing emails* (9.3%), *decreased attention in English contexts* (8.0%), and *difficulty finding similar cases in English on the Internet* (3.1%). The final inter-rater reliability was 0.84, which is considered to be a high agreement. Comments from German, South Korean, and Japanese participants are

indicated with (G), (K), and (J), respectively. The discussion of each concern follows, and design implications based on these concerns will be presented in Section 5.1.

Difficulty Understanding English Email Content. The majority of participant concerns contained anxiety about identifying English phishing emails because the participants struggled to understand the content of English emails. Comments from such participants indicated that the fundamental cause of this concern is their lack of English skills: "*Because my English skills aren't very good, I can't understand English emails at all*"(J) and "*I'm not good at English. Even when I did roughly understand the email, I couldn't grasp its details in English...*"(J). Especially, older participants expressed strong concerns: "*It's been a long time since I learned English, so I can't read it very well*" (K). As a strategy to address this concern, some participants used an online translator. However, they also complained about its inaccuracy: "*I tried to use an online translator, but unfortunately, its translations aren't very good, and they are sometimes very confusing...*" (K). Participants who felt that they could not understand the English emails admitted that they often ignored them: "*Since I can't read English, I usually just ignore English emails*"(J). This is typical security-risk-averse behavior and can certainly prevent English phishing emails, but such biased behavior also creates a risk of opportunity loss by inhibiting communication in English. We conclude that NNESs need support to reduce two distinct risks: English phishing emails and opportunity loss of English communication.

Difficulty Identifying Errors and Unnatural Language in English. Baki et al. [3] found that users generally investigated such language information as writing styles and grammar to identify phishing emails. However, our participants believed that they could not adopt this strategy for emails in English: "*For Japanese emails, I can obviously identify incongruities caused by grammar, nuances, and honorific expressions. However, in English, although I can understand the surface contents of emails, I cannot grasp any language nuances*"(J). Indeed, the result of Section 4.1.3 shows that participants paid less attention to grammatical errors and misspellings in English contexts. This concern is not a simple problem because many participants mentioned not only errors in sentences but subtle unnatural nuances in the language. Participants believed that they needed a high level of English knowledge to overcome this concern: "*Phishing emails are not always obvious. Further English knowledge is necessary to more certainly recognize them*" (G).

Unfamiliarity with English Phishing Emails. Sheng et al. [57] reported that the participants with a high degree of prior exposure to anti-phishing education (i.e., familiarity with phishing) were significantly less susceptible to phishing. However, participants were concerned about their unfamiliarity with English phishing emails:"*... I'm not familiar with the formats and patterns of English phishing emails*" (J) and "*... Compared to Korean ones, English phishing emails are more varied and sneaky, which increases the odds that they will be confusing*" (K). A participant noted the difference in the amount of experience receiving phishing emails written in their native language and those in English as well as the amount that can be learned from familiar media: "*I think that there is a general type of phishing email that is written in Korean. It's an advantage to experience more phishing emails in Korean than similar emails in English. All kinds of media deal with (Korean) phishing emails, so there are more chances to figure out if it's phishing compared to those in English...*" (K).

Decreased Attention in English contexts. Although it is evident that users' attention is essential to identify phishing emails, several participants were concerned that their attention would be reduced in reading English emails: "*... I can't understand the contents of English emails. Thus, I practically panic and worry that I won't make the right decision when I receive it*" (J), and "*Since I can't read English, I blindly open a phishing email to understand the contents*" (J). These concerns reflect security-risk-prone behaviors of NNESs. One participant noted that their attention was decreased because the language of the URLs is English.: "*... If Korean emails provide a link, since its language is different, I might not click on it because it looks different. But for English emails, since the contents are in English and the link is also English, it does not stand out so much. Therefore, I might click on the link more easily than in Korean emails*" (K). It seems to be unique to NNESs to focus on the discrepancy between the languages used in the body of the email and the link (i.e., URL that can use Unicode) respectively, however English emails do not have this feature, suggesting that it is not an effective behavior against English phishing emails.

Difficulty Finding Similar Cases in English on the Internet. In our roleplay task, some participants told us that they used Internet search engines to find similar cases of received suspicious emails in their native language contexts. This means searching on the Internet is an important strategy for identifying phishing emails. However, participants mentioned that they encountered a problem when they searched for similar cases in English: "*When I Google the text of a phishing email in Japanese, I can see if it is phishing by viewing the posted experiences by people who received a similar email. However, for English, I cannot see if it is phishing because it's difficult for me to read the contents of websites from a Google search*" (J); and "*... Even if I can search websites related to the phishing email, it is difficult to determine which information is correct in English contexts*" (J). This matches the findings of Chu et al. [10, 11] who reported that NNESs struggle when viewing and skimming online search results. Although only few participants mentioned this concern (3.1%), we infer that participants who mentioned that they struggled to understand the content of English emails (70.0%) would also have difficulty when searching for similar cases.

On the other hand, the following are three main reasons collected from participants who were confident in identifying

English phishing emails: their attention improves due to a lack of opportunities to receive English emails in their daily life, they can read the elements needed to identify phishing in English, and they believe the mailer and in-house system will detect it. However, regarding the mailer and in-house system, one participant reported that such systems make decisions that lead to lost opportunities: "*Since distinguishing between good and phishing emails is often difficult, you often automatically anticipate phishing or spam in the case of English emails. So sometimes an important email might easily get lost*" (G).

5 Discussion

In a society where native and non-native speakers coexist, it is desirable that there is a minimal discrepancy in communication between native and non-native speakers. The capability of non-native speakers to respond appropriately to emails written in English is a typical example; that is, NNESs are expected to be able to read and understand genuine emails, while correctly ignoring phishing emails even when they are written in English. Through our experiments, we found that NNESs were more prone to engaging in undesirable behaviors when they handled English emails, whether phishing or genuine, and that this tendency varied across countries. We also found that NNESs could not adopt their strategies for identifying phishing emails written in their native languages for English phishing emails. Specifically, they had difficulty in identifying errors and unnatural language and searching similar cases in English contexts. In this section, we first discuss the design implications that aim at supporting NNESs in taking the appropriate action when they need to deal with an email written in English. We then discuss the limitations and future extensions of our study.

5.1 Design Implications

Our findings suggest the need to develop assistive technologies to help NNESs handle English emails correctly. In this section, we present specific design implications (**D1**–**D4**) based on our findings. We also discuss their effectiveness and limitations.

D1: Language-agnostic phishing knowledge base. As a strategy for identifying phishing emails written in their native language, some participants reported that they use Internet search engines to obtain information about similar phishing cases. At the same time, they raised a concern that it would be difficult to take the same approach for identifying English phishing emails. The common challenges derived from our participants' comments (as shown in Section 4.2.2 – Difficulty Finding Similar Cases in English on the Internet) and previous studies of information searches in non-native language [10, 11, 69] are as follows. First, for NNESs, obtaining information in English is a difficult task. Second, even when they find correct information in English, they may not be able to interpret it correctly. Moreover, it is not straightforward for NNESs to ascertain the reliability of information sources. Based on these observations, we propose to develop a *phishing knowledge base*, which (1) is operated by a globally authorized, neutral organization such as an international standardization organization, (2) collects and maintains phishing cases in various languages, and (3) provides *language-agnostic notations* so that NNESs can understand the phishing content. As a previous study on the design of a security indicator implies [19], adopting graphical notations would be effective for solving this problem.

D2: Auto follow-up mechanism Our survey revealed that some NNESs tended to engage in security-risk-averse behavior primarily because the email was written in English. While this behavior may help to reduce the threat of phishing, it could lead to the increase of the risk of losing important opportunities by ignoring all incoming emails written in English. We believe that introducing an auto follow-up mechanism is useful in solving this problem. Gmail [9] has adopted a functionality to provide both an email sender and recipient with a quick reminder that nudges them to follow-up or respond to a potentially important email. For instance, the Gmail inbox displays the message "Received X days ago. Reply?" for the receiver and "Sent X days ago. Follow up?" for the sender.

D3: Training on anti-phishing emails for NNESs Some participants reported that they were confident in identifying phishing emails written in their native language, but it was difficult for them to identify phishing emails written in English because they were unfamiliar with the patterns of English phishing emails. They also reported a concern that their attention was reduced when reading English emails compared to reading their native language emails. One promising strategy to solve such a problem is to provide training. As previous studies have reported, providing a training program is known to be effective in encouraging appropriate action toward received emails, which could contain phishing [35, 57, 58]. Participants in the training program can learn what to watch for in an email and the intrinsic wording to help them identify a phishing message. There are no studies that have shown that important points for identifying phishing emails, e.g., the domain name of a URL in the email, vary greatly across languages. Therefore, it may seem that it is sufficient for non-English speakers to take phishing training in their native language. However, the essential elements in identifying phishing emails are not only the technical points such as domain names in URLs, but also the correct understanding of the content and context of the email, which must be learned through specific examples in English. Therefore, it is desirable for NNESs to participate in English Phishing training. Moreover, because our regression analysis revealed that confidence in English reading increases security-risk-prone behaviors, we believe that English language lessons alone would not be sufficient and that specialized English phishing training for NNESs would be needed. Assessing the effectiveness of such

an educational approach is a challenge for the future.

D4: Machine translation as an assistive tool. Machine translation (MT) services are expected to help NNESs with concerns about their inability to understand English emails. In recent years, the accuracy of MT as well as existing MT services, such as DeepL [16] and Google Translate [23], which are known to generate very natural translations, has improved drastically due to advances in deep learning technology. In fact, many participants mentioned that they relied on MT services when they needed to read emails in English. As it is expected that the quality of MT technology will continue to improve in the future, adoption of MT as an assistive tool could help NNESs identify English phishing emails.

MT is expected to provide the advantages mentioned above; however, the advancement of MT may raise two new concerns: (i) it could interfere with the commonly used phishing identification practice when receiving phishing emails that are likely translated from the original language, i.e., detecting grammatical typos/errors in phishing emails, and (ii) if the phishing email sender uses advanced MT and sends the translated phishing emails written in the recipient's native language, the recipients could be fooled by phishing scams because the emails are written with natural text in their native languages. These observations imply that the strategy that many participants use to fight against phishing emails, i.e., detecting grammatical typos/errors, will no longer be promising in the future. As MT technology improves, strategies for identifying phishing emails that rely solely on grammatical errors should be avoided, and other essential features associated with phishing should be considered.

Because NNESs' concerns in identifying English phishing were not specific to a particular language/culture, we believe the above design implications are generalizable to non-native speakers of other languages.

5.2 Limitations and Future Work

While our survey provides much insight into the challenges faced by NNESs when they receive English emails, there are several limitations.

The purpose of this study was to examine how language barriers impact users' susceptibility to phishing emails. The reason we recruited participants from three countries varying in English proficiency was not to compare cultural effects but to confirm the robustness of our findings. However, in addition to language, cultural differences may have influenced the results of this study. Sawaya et al. [55] and Harbach et al. [26] examined "active" attitudes for secure use of devices or services (e.g., updating software, strengthening passwords, and locking smartphones) and reported that Japanese participants exhibited less secure behavior compared with participants from other countries. We cannot conclude that those results are inconsistent with our result indicating that Japanese participants are less likely to engage in security-risk-prone behavior, as shown in Table 4. People's behavior may vary depending on the context, thus our work focused on revealing behavioral tendencies in the context of phishing email. Furthermore, demographic differences may have influenced the results of this study. The high proportion of IT professionals among German participants may have influenced our survey results, although our regression analysis revealed that IT expertise did not significantly affect participants' security-risk-prone/averse behaviors. The different types of prior training provided in each country also may have influenced participants' behaviors. It is complicated to conduct a survey of susceptibility to phishing emails that completely separates the effects of language from the effects of cultural and demographic differences. To reduce such effects, we adopted a between-subjects design for each country instead of directly comparing participants' results per country in our roleplay task.

Through our user study, we tested whether participants were willing to follow the instructions (i.e., clicking on the links or opening the attachment files) in the phishing emails. However, after accessing a website in an actual phishing attack, users may see an alerting security indicator and realize that the website is a phishing website, and the attack may not be successful. This study did not take such cases into account. To determine the likelihood of NNESs falling victim to a phishing attack, it is necessary to observe the overall decision-making process of NNESs after reading a phishing email written in English and visiting a website. Conducting user studies with more strict ecological validity is a challenge for future research. In addition, further research is needed to investigate NNESs' behavior when they receive more sophisticated spear-phishing emails that are highly aligned with their personal contexts.

6 Conclusion

Through our scenario-based roleplay task, we showed how non-native English speakers (NNESs) adopted security-risk-prone/averse strategies toward emails in their native language and English. Specifically, we found that participants adopted more security-risk-averse behaviors (i.e., ignoring emails without careful inspection) when the emails were written in English rather than in their native languages. In addition, our qualitative analysis of their open-ended answers revealed five main factors that formed their concerns for identifying English phishing emails; these include difficulty identifying language errors, difficulty finding similar cases, and unfamiliarity. Our findings bring the unique insight that NNESs may have different concerns and strategies for avoiding phishing emails. It indicates the importance of considering language barriers when designing interventions to support people in combating phishing attacks. Implementing specific anti-phishing interventions for NNESs based on our findings is an important research effort to reduce communication difficulties between native and non-native English speakers.

References

[1] Sara Albakry, Kami Vaniea, and Maria K. Wolters. What is this url's destination? empirical evaluation of users' url reading. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI'20, 2020.

[2] Ibrahim Mohammed Alseadoon, Rabie A Ramadan, and Ahmed Y Khedr. Cultural impact on users' ability to protect themselves against phishing websites. *International Journal of Computer Science and Network Security*, 17(11), 2017.

[3] Shahryar Baki, Rakesh Verma, Arjun Mukherjee, and Omprakash Gnawali. Scaling and effectiveness of email masquerade attacks: Exploiting natural language generation. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, AsiaCCS'17, 2017.

[4] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[5] Marcus Butavicius, Kathryn Parsons, Malcolm Pattinson, and Agata McCormac. Breaching the human firewall: Social engineering in phishing and spear-phishing emails. In *Proceedings of the 26th Australasian Conference on Information Systems*, ACIS'15, 2015.

[6] Marcus A Butavicius, Kathryn Parsons, Malcolm R Pattinson, Agata McCormac, Dragana Calic, and Meredith Lillie. Understanding susceptibility to phishing emails: Assessing the impact of individual differences and culture. In *Proceedings of the 11th International Symposium on Human Aspects of Information Security & Assurance*, HAISA'17, 2017.

[7] Casey Canfield, Alex Davis, Baruch Fischhoff, Alain Forget, Sarah Pearman, and Jeremy Thomas. Replication: Challenges in using data logs to validate phishing detection ability metrics. In *Proceedings of the 13th Symposium on Usable Privacy and Security*, SOUPS'17, 2017.

[8] Casey Inez Canfield, Baruch Fischhoff, and Alex Davis. Quantifying phishing susceptibility for detection and behavior decisions. *Human factors*, 58(8):1158–1172, 2016.

[9] G Suite Learning Center. 7.3 remember to follow up. https://support.google.com/a/users/answer/9259771#7.3, 2020 (accessed September 17, 2020).

[10] Peng Chu, Eszter Jozsa, Anita Komlodi, and Karoly Hercegfi. An exploratory study on search behavior in different languages. In *Proceedings of the 4th Information Interaction in Context Symposium*, IIIX'20, 2012.

[11] Peng Chu, Anita Komlodi, and Gyöngyi Rózsa. Online search in english as a non-native language. *The Association for Information Science and Technology*, 52(1):1–9, 2015.

[12] Robert B Cialdini. *Influence: The psychology of persuasion*. 1984.

[13] Albert Costa, Alice Foucart, Inbal Arnon, Melina Aparici, and Jose Apesteguia. "piensa" twice: On the foreign language effect in decision making. *Cognition*, 130:236–254, 2013.

[14] Robert E Crossler, Allen C Johnston, Paul Benjamin Lowry, Qing Hu, Merrill Warkentin, and Richard Baskerville. Future directions for behavioral information security research. *computers & security*, 32:90–101, 2013.

[15] Avisha Das, Shahryar Baki, Ayman El Aassal, Rakesh Verma, and Arthur Dunbar. Sok: A comprehensive reexamination of phishing research from the security perspective. *IEEE Communications Surveys & Tutorials*, 22(1):671–708, 2020.

[16] DeepL. Deepl translate. https://www.deepl.com/translator, 2020 (accessed September 17, 2020).

[17] Rachna Dhamija, J Doug Tygar, and Marti Hearst. Why phishing works. In *Proceedings of the 2006 SIGCHI conference on Human Factors in computing systems*, CHI'06, 2006.

[18] Julie S Downs, Mandy B Holbrook, and Lorrie Faith Cranor. Decision strategies and susceptibility to phishing. In *Proceedings of the 2nd Symposium on Usable Privacy and Security*, SOUPS'06, 2006.

[19] Adrienne Porter Felt, Robert W. Reeder, Alex Ainslie, Helen Harris, Max Walker, Christopher Thompson, Mustafa Embre Acer, Elisabeth Morant, and Sunny Consolvo. Rethinking connection security indicators. In *Proceedings of the 20th Symposium on Usable Privacy and Security*, SOUPS'16, 2016.

[20] EF Education First. Ef english proficiency index. https://www.ef.com/wwen/epi/, 2019 (accessed June 13, 2020).

[21] Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23, 1981.

[22] Waldo Rocha Flores, Hannes Holm, Marcus Nohlberg, and Mathias Ekstedt. Investigating personal determinants of phishing and the effect of national culture. *Information & Computer Security*, 23:178–199, 2015.

[23] Google. Google translate. https://translate.google.com/, 2020 (accessed September 17, 2020).

[24] Kristen K Greene, Michelle P Steves, Mary F Theofanos, and Jennifer Kostick. User context: an explanatory variable in phishing susceptibility. In *Proceedings of the 2018 Workshop on Usable Security*, USEC'18, 2018.

[25] Tzipora Halevi, Nasir Memon, and Oded Nov. Spear-phishing in the wild: A real-world study of personality, phishing self-efficacy and vulnerability to spear-phishing attacks. *SSRN Electronic Journal*, 2015.

[26] Marian Harbach, Alexander De Luca, Nathan Malkin, and Serge Egelman. Keep on lockin'in the free world: A multi-national comparison of smartphone locking. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI'16, 2016.

[27] Brynne Harrison, Elena Svetieva, and Arun Vishwanath. Individual processing of phishing emails. *Online Information Review*, 40(2):265–281, 2016.

[28] Farkhondeh Hassandoust, Harminder Singh, and Jocelyn E Williams. How contextualisation affects the vulnerability of individuals to phishing attempts. In *Proceedings of the 23th Pacific Asia Conference on Information Systems*, PACIS'19, 2019.

[29] Geert Hofstede. National cultures in four dimensions: A research-based theory of cultural differences among nations. *International Studies of Management & Organization*, 13(1-2):46–74, 1983.

[30] Hannes Holm, Waldo Rocha Flores, Marcus Nohlberg, and Mathias Ekstedt. An empirical investigation of the effect of target-related information in phishing attacks. In *Proceedings of the 11th IEEE International Enterprise Distributed Object Computing Conference Workshops and Demonstrations*, 2014.

[31] Kyung Wha Hong, Christopher Kelley, Rucha Tembe, Emerson Murphy-Hill, and Christopher Mayhorn. Keeping up with the joneses: Assessing phishing susceptibility in an email task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 57:1012–1016, 2013.

[32] Tom N Jagatic, Nathaniel A Johnson, Markus Jakobsson, and Filippo Menczer. Social phishing. *Communications of the ACM*, 50(10):94–100, 2007.

[33] Joakim Kävrestad, Rickard Pettersson, and Marcus Nohlberg. The language effect in phishing susceptibility. In *Proceedings of the 6th International Workshop on Socio-Technical Perspective in IS Development*, STPIS'20, 2020.

[34] Boaz Keysar, Sayuri L Hayakawa, and Sun Gyu An. The foreign-language effect: Thinking in a foreign tongue reduces decision biases. *Psychological science*, 23(6):661–668, 2012.

[35] Ponnurangam Kumaraguru, Justin Cranshaw, Alessandro Acquisti, Lorrie Cranor, Jason Hong, Mary Ann Blair, and Theodore Pham. School of phish: A real-world evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, SOUPS'09, 2009.

[36] Harjinder Singh Lallie, Lynsay A Shepherd, Jason RC Nurse, Arnau Erola, Gregory Epiphaniou, Carsten Maple, and Xavier Bellekens. Cyber security in the age of covid-19: A timeline and analysis of cyber-crime and cyber-attacks during the pandemic. *arXiv preprint arXiv:2006.11929*, 2020.

[37] Elmer Lastdrager, Inés Carvajal Gallardo, Pieter Hartel, and Marianne Junger. How effective is anti-phishing training for children? In *Proceedings of 13th Symposium on Usable Privacy and Security*, SOUPS'17, 2017.

[38] Wanru Li, James Lee, Justin Purl, Frank Greitzer, Bahram Yousefi, and Kathryn Laskey. Experimental investigation of demographic factors related to phishing susceptibility. In *Proceedings of the 53rd Hawaii International Conference on System Sciences*, HICCS'20, 2020.

[39] Tian Lin, Daniel E Capecci, Donovan M Ellis, Harold A Rocha, Sandeep Dommaraju, Daniela S Oliveira, and Natalie C Ebner. Susceptibility to spear-phishing emails: Effects of internet user demographics and email content. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5):1–28, 2019.

[40] Macromill. Macromill. https://group.macromill.com/, 2000 (accessed June 13, 2020).

[41] MillerSmiles.co.uk. Millersmiles.co.uk. http://www.millersmiles.co.uk/, 2003 (accessed June 18, 2020).

[42] Gregory D Moody, Dennis F Galletta, and Brian Kimball Dunn. Which phish get caught? an exploratory study of individuals' susceptibility to phishing. *European Journal of Information Systems*, 26(6):564–584, 2017.

[43] Keika Mori, Takuya Watanabe, Yunao Zhou, Ayako Akiyama Hasegawa, Mitsuaki Akiyama, and Tatsuya Mori. Comparative analysis of three language spheres: Are linguistic and cultural differences reflected in password selection habits? In *Proceedings of the 2019 IEEE European Workshop on Usable Security*, EuroUSEC'19, 2019.

[44] Mattia Mossano, Kami Vaniea, Lukas Aldag, Reyhan Düzgün, Peter Mayer, and Melanie Volkamer. Analysis of publicly available anti-phishing webpages: contradicting information, lack of concrete advice and very narrow attack vector. In *Proceedings of the 5th European Workshop on Usable Security*, EuroUSEC'20, 2020.

[45] Tsedal Neeley. Language matters: Status loss and achieved status distinctions in global organizations. *Journal of Organization Science*, 24(2):476–497, 2013.

[46] James Nicholson, Lynne Coventry, and Pam Briggs. Can we fight social engineering attacks by social means? assessing social salience as a means to improve phish detection. In *Proceedings of the 13th Symposium on Usable Privacy and Security*, SOUPS'17, 2017.

[47] Daniela Oliveira, Harold Rocha, Huizi Yang, Donovan Ellis, Sandeep Dommaraju, Melis Muradoglu, Devon Weir, Adam Soliman, Tian Lin, and Natalie Ebner. Dissecting spear phishing emails for older vs young adults: On the interplay of weapons of influence and life domains in predicting susceptibility to phishing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI'17, 2017.

[48] Kathryn Parsons, Marcus Butavicius, Malcolm Pattinson, Dragana Calic, Agata Mccormac, and Cate Jerram. Do users focus on the correct cues to differentiate between phishing and genuine emails? In *Proceedings of the 26th Australasian Conference on Information Systems*, ACIS'15, 2015.

[49] Kathryn Parsons, Agata McCormac, Malcolm Pattinson, Marcus Butavicius, and Cate Jerram. Phishing for the truth: A scenario-based experiment of users' behavioural response to emails. In *Proceedings of the 28th IFIP TC 11 International Information Security and Privacy Conference*, IFIP SEC'13, 2013.

[50] Prashanth Rajivan and Cleotilde Gonzalez. Creative persuasion: A study on adversarial behaviors and strategies in phishing attacks. *Frontiers in psychology*, 9:135, 2018.

[51] David Rear. The language deficit: a comparison of the critical thinking skills of asian students in first and second language contexts. *Asian-Pacific Journal of Second and Foreign Language Education*, 2(13), 2017.

[52] Elissa M Redmiles. "should i worry" a cross-cultural examination of account security incident response. In *Proceedings of the 2019 IEEE Symposium on Security and Privacy*, S&P'19, 2019.

[53] Joshua Reynolds, Deepak Kumar, Zane Ma, Rohan Subramanian, Meishan Wu, Martin Shelton, Joshua Mason, Emily Stark, and Michael Bailey. Measuring identity confusion with uniform resource locators. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI'20, 2020.

[54] Stephen A Ross. Some stronger measures of risk aversion in the small and the large with applications. *Econometrica: Journal of the Econometric Society*, pages 621–638, 1981.

[55] Yukiko Sawaya, Mahmood Sharif, Nicolas Christin, Ayumu Kubota, Akihiro Nakarai, and Akira Yamada. Self-confidence trumps knowledge: A cross-cultural study of security behavior. In *Proceedings the 2017 CHI Conference on Human Factors in Computing Systems*, CHI'17, 2017.

[56] Anjum N Shaikh, Antesar M Shabut, and MA Hossain. A literature review on phishing crime, prevention review and investigation of gaps. In *Proceedings of the 10th International Conference on Software, Knowledge, Information Management & Applications*, SKIMA'16, 2016.

[57] Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions. In *Proceedings of the 2010 SIGCHI Conference on Human Factors in Computing Systems*, CHI'10, 2010.

[58] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Anti-phishing phil: The design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the 3rd Symposium on Usable Privacy and Security*, SOUPS'07, 2007.

[59] Michelle P Steves, Kristen K Greene, and Mary F Theofanos. A phish scale: Rating human phishing message detection difficulty. In *Proceedings of the 2019 Workshop on Usable Security*, USEC'19, 2019.

[60] World Wide Web Technology Surveys. Usage statistics of content languages for websites. https://w3techs.com/technologies/overview/content_language, 2021 (accessed February 25, 2021).

[61] Yohtaro Takano and Akiko Noda. A temporary decline of thinking ability during foreign language processing.

Journal of Cross-Cultural Psychology, 24(4):445–462, 1993.

[62] Rucha Tembe, Olga Zielinska, Yuqi Liu, Kyung Wha Hong, Emerson Murphy-Hill, Chris Mayhorn, and Xi Ge. Phishing in international waters: exploring cross-national differences in phishing conceptualizations between chinese, indian and american samples. In *Proceedings of the 2014 Symposium and Bootcamp on the Science of Security*, HotSoS'14, 2014.

[63] Amber Van Der Heijden and Luca Allodi. Cognitive triaging of phishing attacks. In *Proceedings of the 28th USENIX Security Symposium*, SEC'19, 2019.

[64] Verizon. 2020 data breach investigations report. https://enterprise.verizon.com/resources/reports/dbir/, 2020 (accessed June 17, 2020).

[65] Arun Vishwanath, Tejaswini Herath, Rui Chen, Jingguo Wang, and H Raghav Rao. Why do people get phished? testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems*, 51(3):576–586, 2011.

[66] Stefan Volk, Tine Köhler, and Markus Pudelko. Brain drain: The cognitive neuroscience of foreign language processing in multinational corporations. *Journal of International Business Studies*, 45(7):862–885, 2014.

[67] Jingguo Wang, Tejaswini Herath, Rui Chen, Arun Vishwanath, and H Raghav Rao. Phishing susceptibility: An investigation into the processing of a targeted spear phishing email. *IEEE transactions on professional communication*, 55(4):345–362, 2012.

[68] Emma J Williams, Joanne Hinds, and Adam N Joinson. Exploring susceptibility to phishing in the workplace. *International Journal of Human-Computer Studies*, 120:1–13, 2018.

[69] Alyson L. Young, Anita Komlodi, Gyöngyi Rózsa, and Peng Chub. Evaluating the credibility of english web sources as a foreign-language searcher. *The Association for Information Science and Technology*, 53(1):1–9, 2016.

A Questionnaire

Each participant read and answered the questionnaire in their native language. Participants were randomly assigned to a group where they were shown emails written in their native language or English. Asterisk (*) indicates that the sentences were arranged by the participant's country: Germany, South Korea, or Japan. Double asterisks (**) indicate that the sentences were dynamically arranged according to the participant's preceding answers.

Screening survey

Survey Title: Survey of emails written in German/Korean/Japanese* or English.

Number of questions: 5 in the screening survey and 17 in the main survey (time required: about 25 minutes).

Participation compensation: 4 EUR / 5500 KRW / 500 JPY (for those who participated in the main survey)

Data handling: This questionnaire is conducted anonymously. Responses to it will be used for academic research. The aggregated results of the answers to the multiple choice questions will be published in an academic journal, and the answers to the open-ended questions may be published in an academic journal with a non-personally identifiable form. The answers will be provided to requesting organizations, and translations may be outsourced to a third party. The answers will be protected as confidential information.

Note: This survey has several open-ended questions. To help us improve the quality of our research, please be as specific as possible about your opinions. This survey also includes several image-based questions.

I agree with the above information and agree to participate in this survey.

- ○ Yes, I agree with the above statement and I will participate in this survey.
- ○ No.

Q01. How old are you?

- ○ 18-29 years old
- ○ 30-39 years old
- ○ 40-49 years old
- ○ 50-59 years old
- ○ 60 years or older
- ○ Prefer not to answer

Q02. What is your gender (self-identified gender)?

- ○ Male
- ○ Female
- ○ Other
- ○ Prefer not to answer

Q03. Which of the following best describes your current occupational status? Please select the most applicable answer.

- ○ Work (full-time)
- ○ Work (part-time)
- ○ Student
- ○ Unemployed or retired (including homemaker)

Q04. This question is designed to verify that you have read the question carefully.
Please select both “No” and “Other”.

- ☐ Yes
- ☐ No
- ☐ Other
- ☐ Prefer not to answer

Q05. What is your native language (the language you primarily spoke before you were 10 years old)?

- ○ German/Korea/Japanese*
- ○ English
- ○ Other

Main survey

Q01. Are you an expert in the fields of information technology (IT), computer engineering, or computer science?

- ○ Yes
- ○ No

Q02. Which of the following best describes your highest achieved education level? Please select the most applicable answer.

- ○ Some high school
- ○ High school graduate
- ○ Some college, no degree
- ○ Associate’s degree
- ○ Bachelor’s degree
- ○ Graduate degree
- ○ Other
- ○ Prefer not to answer

Q03. How confident are you in your ability to read English? Please select the most applicable answer.

- ○ Very unconfident
- ○ Unconfident
- ○ Somewhat unconfident
- ○ Somewhat confident
- ○ Confident
- ○ Very confident

Q04. How many years have you studied English in total, including self-study? Please round your answer down to the nearest whole number.
( ) years

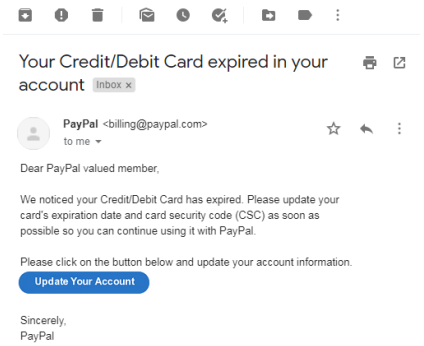
Q05. How many emails (both work-related and personal) do you receive on average on a typical weekday? Please include auto-send emails. If you receive less than one email a day on average, answer "0".
Emails written in German/Korean/Japanese*: ( )
Emails written in English: ( )

From here, you will answer by looking at email screenshots. Answer by looking at the screenshots without actually searching or accessing the information in them. The recipient of the following email is Max Mustermann / Hong Gil-dong / Taro Yamada*. Please answer questions 6-9 as if you were Max Mustermann / Hong Gil-dong / Taro Yamada*.

Profile of Max Mustermann / Hong Gil-dong / Taro Yamada*

- Name: Mr. Max Mustermann / Hong Gil-dong / Taro Yamada*
- Country of residence: Germany / South Korea / Japan*
- Occupation: office worker
- Employer: ABC Company
- Boss: Erika Müller(erika.mueller@abccompany.com) / Hong Gil-soon (gilsoon.hong@abccompany.com) / Hanako Tanaka (hanako.tanaka@abccompany.com)*
- Email address of IT service department: it-service@abccompany.com
- Online services he uses:
  - PayPal
    - Online payments service. He uses this service for private online shopping. He registered his private email address and his credit card information with this service.
  - LinkedIn
    - Online networking services. He uses this service to build his network. He registered his private email address with this service.
  - Zoom
    - Video conferencing service. He uses this service for working from home. He registered his business email address with this service.

Email (a): An email sent to a private email address.

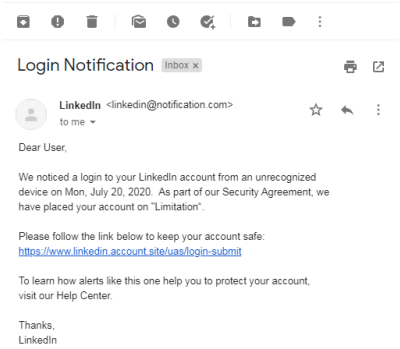


Q06. How would you respond if you received email (a)?

- ○ I’d ignore it without referring to any other information than this screenshot.

- ○ I'd follow its instruction without referring to any other information than this screenshot.
- ○ I'd refer to some other information than this screenshot to decide how to respond.
  Please specify. ( )
- ○ Other
  Please specify. ( )

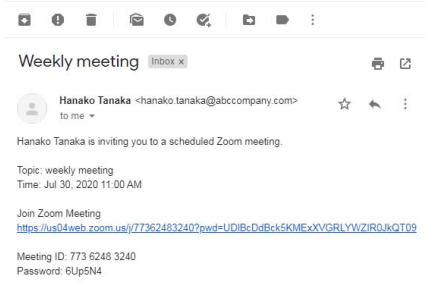
Email (b): An email sent to a private email address.



Q07. How would you respond if you received email (b)?

- ○ I'd ignore it without referring to any other information than this screenshot.
- ○ I'd follow its instruction without referring to any other information than this screenshot.
- ○ I'd refer to some other information than this screenshot to decide how to respond.
  Please specify. ( )
- ○ Other
  Please specify. ( )

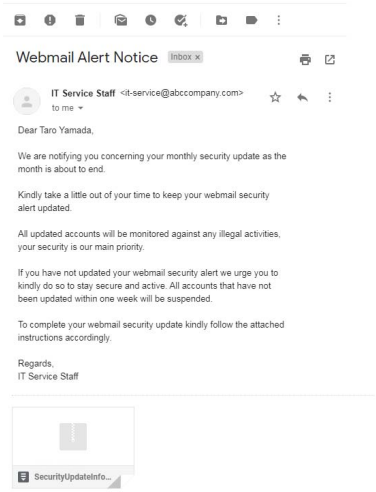
Email (c): An email sent to a business email address.



Q08. How would you respond if you received email (c)?

- ○ I'd ignore it without referring to any other information than this screenshot.
- ○ I'd follow its instruction without referring to any other information than this screenshot.
- ○ I'd refer to some other information than this screenshot to decide how to respond.
  Please specify. ( )
- ○ Other
  Please specify. ( )

Email (d): An email sent to a business email address.



Q09. How would you respond if you received email (d)?

- ○ I'd ignore it without referring to any other information than this screenshot.
- ○ I'd follow its instruction without referring to any other information than this screenshot.
- ○ I'd refer to some other information than this screenshot to decide how to respond.
  Please specify. ( )
- ○ Other
  Please specify. ( )

Q10. When you receive an email written in English/German/Korean/Japanese*[3], to which parts do you usually pay attention? Choose the three items from the list below to which you pay the most attention.

- ☐ Sender's name
- ☐ Sender's email address
- ☐ Subject line
- ☐ Grammatical errors/misspellings in the language
- ☐ Formality of language
- ☐ Urgency of message
- ☐ Your involvement with the email
- ☐ URLs
- ☐ Attachment file

Q11. This question is designed to verify that you have carefully read the question.
Please select both "No" and "Prefer not to answer".

[3] We displayed the language which corresponded with the emails that the participant was presented in Q06-Q09.

- ☐ Yes
- ☐ No
- ☐ Other
- ☐ Prefer not to answer

From here, we will ask about your experience and perception of phishing. Phishing is online fraud that acquires sensitive information primarily by masquerading as a legitimate business or a reputable person.

Q12. How often do you receive both work-related and personal emails that are assumed to be phishing? Do not include phishing-training emails from your company. Please select the most applicable answer.
Phishing emails written in German/Korean/Japanese*:

- ○ Less than once a year
- ○ Once a year
- ○ Once every few months
- ○ Once a month
- ○ Once a week
- ○ Once a day
- ○ More than once a day
- ○ I don't know

Phishing emails written in English:

- ○ Less than once a year
- ○ Once a year
- ○ Once every few months
- ○ Once a month
- ○ Once a week
- ○ Once a day
- ○ More than once a day
- ○ I don't know

Q13. Have you ever been deceived by a phishing email? Being deceived by a phishing email means that you visited a website linked in the phishing email or opened a file attached to the phishing email, regardless whether you were directly damaged. Please answer the total number of work-related and personal experiences. Do not include your experience with phishing-training emails from your company.
Experience with phishing emails written in German/Korean/Japanese*:

- ○ I have been deceived
  Approximately ( ) times
- ○ I have been deceived, but I don't remember how many times
- ○ I have never been deceived
- ○ I don't know

Experience with phishing emails written in English:

- ○ I have been deceived
  Approximately ( ) times
- ○ I have been deceived, but I don't remember how many times
- ○ I have never been deceived
- ○ I don't know

An optional question for participants who answered "I have been deceived" or "I have been deceived, but I don't remember how many times" in Q13-2.
Q14**. If you remember the content of the phishing email in English, describe it as specifically as possible (e.g., the company/service/person the attacker masqueraded, the purpose and its requests, and why you were unable to identify it as a phishing email). If you have been deceived more than once, please tell us about the most recent phishing email you received.

Q15. This question is designed to verify that you are carefully reading the question. Please choose one of the following statements that fits the definition of phishing:

- ○ An attacker encrypts files on your device
- ○ An attacker masquerades as a legitimate company/service/person and asks for sensitive information
- ○ An attacker sends a massive amount of traffic to a target website to disable it.

Q16. To what extent do you agree with the following statements? Please select the most applicable answer.
"I can always identify a phishing email written in German/Korean/Japanese*."

- ○ Strongly disagree
- ○ Disagree
- ○ Somewhat disagree
- ○ Somewhat agree
- ○ Agree
- ○ Strongly agree

"I can always identify phishing email written in English."

- ○ Strongly disagree
- ○ Disagree
- ○ Somewhat disagree
- ○ Somewhat agree
- ○ Agree
- ○ Strongly agree

Q17. Please specify why you think you can/cannot** identify a phishing email written in English. Answer by comparing it with the German/Korean/Japanese* case.

B Results of the Roleplay Task

Table 6: Detailed results of the roleplay task

			N	% Ignore	% Follow	Other				
						% Check the website without a link	% Ask a related person	% Internet search engine	% Online translator	% Other
(a)	Native	Germany	140	48.6%	27.1%	17.1%	0.0%	2.1%	0.0%	5.0%
		South Korea	141	20.6%	41.8%	22.0%	0.0%	8.5%	0.0%	7.1%
		Japan	148	44.6%	8.8%	26.4%	0.0%	15.5%	0.0%	4.7%
	English	Germany	144	61.8%	22.2%	10.4%	0.0%	1.4%	0.0%	4.2%
		South Korea	135	41.5%	25.9%	21.5%	0.0%	2.2%	5.2%	3.7%
		Japan	154	56.5%	14.3%	11.7%	0.6%	7.8%	6.5%	2.6%
(b)	Native	Germany	140	50.7%	30.7%	10.0%	0.0%	2.1%	0.0%	6.4%
		South Korea	141	29.8%	43.3%	11.3%	0.0%	7.1%	0.0%	7.8%
		Japan	148	55.4%	8.8%	19.6%	0.0%	11.5%	0.0%	5.4%
	English	Germany	144	61.8%	23.6%	8.3%	0.0%	0.0%	0.0%	6.3%
		South Korea	135	46.7%	25.2%	8.9%	0.0%	3.7%	5.9%	8.9%
		Japan	154	57.8%	14.9%	9.7%	0.6%	5.8%	5.2%	5.8%
(c)	Native	Germany	140	19.3%	68.6%	4.3%	5.0%	0.0%	0.0%	2.9%
		South Korea	141	21.3%	57.4%	3.5%	9.9%	0.7%	0.0%	7.1%
		Japan	148	18.9%	56.1%	3.4%	12.8%	0.0%	0.0%	8.8%
	English	Germany	144	24.3%	63.9%	2.1%	5.6%	0.0%	0.0%	4.2%
		South Korea	135	28.9%	43.7%	3.0%	11.9%	0.0%	2.2%	10.4%
		Japan	154	29.2%	43.5%	5.2%	8.4%	0.0%	5.2%	8.4%
(d)	Native	Germany	140	32.1%	53.6%	0.0%	11.4%	0.7%	0.0%	2.1%
		South Korea	141	30.5%	50.4%	0.0%	10.6%	1.4%	0.0%	7.1%
		Japan	148	46.6%	23.6%	0.0%	16.9%	4.1%	0.0%	8.8%
	English	Germany	144	40.3%	45.8%	0.0%	7.6%	0.0%	1.4%	4.9%
		South Korea	135	37.0%	31.1%	0.0%	13.3%	0.7%	5.9%	11.9%
		Japan	154	48.7%	24.7%	0.0%	9.1%	1.3%	9.1%	7.8%

SoK: Still Plenty of Phish in the Sea — A Taxonomy of User-Oriented Phishing Interventions and Avenues for Future Research

Anjuli Franz, Verena Zimmermann, Gregor Albrecht, Katrin Hartwig,
Christian Reuter, Alexander Benlian, Joachim Vogt
Technical University of Darmstadt

Abstract

Phishing is a prevalent cyber threat, targeting individuals and organizations alike. Previous approaches on anti-phishing measures have started to recognize the role of the user, who, at the center of the target, builds the last line of defense. However, user-oriented phishing interventions are fragmented across a diverse research landscape, which has not been systematized to date. This makes it challenging to gain an overview of the various approaches taken by prior works.

In this paper, we present a taxonomy of phishing interventions based on a systematic literature analysis. We shed light on the diversity of existing approaches by analyzing them with respect to the intervention type, the addressed phishing attack vector, the time at which the intervention takes place, and the required user interaction. Furthermore, we highlight shortcomings and challenges emerging from both our literature sample and prior meta-analyses, and discuss them in the light of current movements in the field of usable security. With this article, we hope to provide useful directions for future works on phishing interventions.

1 Introduction

Phishing is a frequently employed cyber attack to get hold of users' sensitive information, such as login details or banking account numbers. Furthermore, criminals increasingly use phishing attacks to distribute malware [90]. The consequences of a successful attack can reach from individual personal losses or compromised accounts to complete organizations or networks being infected by malware, often combined with ransom demands. For example, the years between 2014 and 2020 were marked by *Emotet*, a modular trojan using targeted phishing emails with weaponized Microsoft Word files [27, 58]. It is crucial to consider that phishing attacks do not primarily target hardware or software vulnerabilities, but the user – the human factor within the socio-technical system. While there are several tools and approaches that aim to identify malicious contents automatically (e.g., [78, 82]), the increasingly sophisticated and personalized nature of phishing attacks makes it hard for algorithms to detect and block phishing emails, websites, or malicious software. This leaves a large amount of responsibility to the user. However, detecting phishing attempts is not the user's first priority [93], for instance, while using email programs: Instead, users in various contexts aim to efficiently solve their tasks and answer what they perceive to be emails sent by customers or colleagues when they become victims of a phishing attack.

To enable users to be the ultimate wall of defense in cyber security, research and practice have developed a number of user-oriented interventions against phishing attacks. Among those are education and training approaches (e.g., [12, 44, 72]), where users develop knowledge and skills that they can transfer to real-world phishing attempts. To complement these, awareness-raising measures or design considerations (e.g., [25, 51, 56, 61]) aim to guide users towards secure online behavior in situ.

While developing adequate countermeasures that assist end-users in combating phishing attacks is highly relevant, finding both effective and usable user-oriented phishing interventions is still an unresolved problem [3]. Considering the diverse research landscape on phishing interventions across various research disciplines (e.g., cyber security, human-computer interaction, or social science), it is challenging to gain an overview of what types of interventions have already been investigated. The design of interventions may significantly differ between phishing attack vectors, the moment at which the intervention takes place, or approaches that increase the attention in a specific moment vs. those that encourage long-term capability to deal with phishing attacks autonomously.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

To our knowledge, a comprehensive literature review of existing approaches is missing to date. We argue that a systematization of prior phishing interventions, particularly with respect to their variety across multiple characteristics, will help to identify trends and gaps in the phishing intervention literature. Furthermore, a discussion in the light of current usable security movements will lead to a better understanding of promising directions for successful user assistance in the phishing context. Our research thus aims to shed light on the following two research questions:

RQ1: How does current research on user-oriented phishing interventions tackle the aim of guiding users towards secure online behavior?

RQ2: Which avenues for future research emerge from the existing phishing intervention literature?

In this work, we offer a comprehensive systematization of user-oriented phishing interventions with respect to the intervention type, the addressed attack vector, the moment at which the intervention takes place, as well as the degree of user interaction. We thereby complement broader reviews such as the work of Zhang-Kennedy & Chiasson [99], who have reviewed tools for cyber security awareness and education more generally. Our contributions are threefold: First, we present an extensive literature analysis of prior research on user-oriented phishing interventions [9, 69], bridging the research streams of both educational and design measures. Guided by previous rudiments of phishing intervention classifications [39, 43, 85, 94], we introduce a novel taxonomy of user-oriented phishing interventions consisting of four categories and ten subcategories. Second, we explore central characteristics such as the time at which the intervention takes place throughout the user's decision process, which phishing attack vectors are commonly addressed by the studied interventions, and the degree of user interaction required. Beyond that, we thirdly take into account critical considerations of leading usable security researchers (e.g., [20, 67, 84]) and discuss shortcomings of prior phishing intervention approaches. In summary, we offer a novel insight into phishing intervention research and present potential avenues for future works.

2 Methodology

To categorize and understand the landscape of existing phishing interventions, we have conducted a systematic literature review, following the "preferred reporting items for a systematic review and meta-analysis" (PRISMA) guideline [53, 55]. Literature reviews have been argued to play an important role in developing domain knowledge, e.g., by synthesizing prior research works, identifying research gaps, and developing a research agenda [69]. To cover the diverse research landscape, our initial search comprised the databases ACM Digital Library, IEEE Xplore, and Web of Science. The search was limited to peer-reviewed studies in English that were available as of June 2020.

The search term was identical across databases and applied to the title and abstract of all included articles. For an article to be included in the analysis, it had to contain the term *phish** and one of the following terms to allow for a plurality of intervention types: *interven** OR *prevent** OR *educat** OR *detect** OR *train** OR *nudg** OR *appeal*.

In addition to the database search, we analyzed the Google Scholar top ten security conferences and journals as well as the A* and A CORE-ranked security conferences and journals. Most of them had already been included in the analyzed databases (e.g., CHI, S&P, CCS, Computers & Security). Only journals and conferences that had not been covered by the previous database search underwent an additional manual title search. These included the USENIX Network and Distributed System Security Symposium NDSS and the accompanying usable security events USEC and EuroUSEC, as well as the USENIX Security Symposium and the co-located SOUPS conference from 2014 onwards[1]. In addition to our search term-based search, we have complemented our sample with two other relevant articles that we became aware of through our literature research.

With the above-described search procedure, we have identified a total of 2,124 publications. Afterward, we have conducted a title and abstract screening to exclude irrelevant articles. Articles were excluded if they matched one of the following criteria:

- Deals with a different topic not related to phishing in the sense of cyber security
- Intervention is not user-oriented in that the user cannot see or act upon an intervention (e.g., an algorithm that invisibly filters and blocks suspicious emails)

Table 1 in the appendix details the distribution across the different databases before and after the title and abstract screening. After the aforementioned procedure as well as the deletion of two duplicates, a total of 80 articles remained for a detailed analysis. As for the full-text screening, we have read and analyzed the 80 articles independently among the authors to ensure best possible thoroughness. Since this literature review has emerged from a cross-disciplinary collaboration between seven security researchers with backgrounds in computer science, information systems and psychology, we were able to discuss the literature from various angles and finally agreed on one final review. The full-text analysis further reduced the literature count by 16 articles: First, we excluded research works that did not address a user-oriented phishing intervention in the full text (see second exclusion criterion above). Second, we excluded similar articles by the same authors (e.g., a conference paper and a subsequent, very similar journal publication), and kept only the latest and more extensive version. Our final literature sample thus includes 64 articles.

[1] Before 2014, the SOUPS proceedings were included in the ACM database.

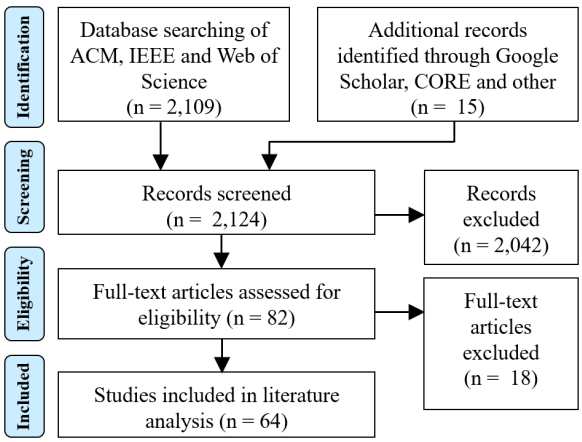


Figure 1: PRISMA diagram of literature screening process.

Figure 1 shows a flowchart that details the number of screened, excluded and included articles following the PRISMA statement [53, 55].

3 Results

In the following, we present a detailed analysis of our literature sample. We first provide an overview of the **methodological range** employed by previous phishing intervention research (Section 3.1). Categorizing the studied interventions with regard to their design and intended effect, we then derive a **taxonomy of user-oriented phishing interventions** (Section 3.2). We further consider the **phishing attack vector** that the intervention aims to address (Section 3.3), the **time at which the intervention takes place** (Section 3.4), as well as the **degree of user interaction** (Section 3.5).

For a comprehensive categorization of the analyzed phishing interventions across the whole literature sample, please refer to Table 2 in the appendix.

3.1 Overview of Methodological Approaches

With respect to the methodological approach, 13 research works have presented exclusively **conceptual ideas** of phishing interventions. For example, Dhamija & Tygar [24] have discussed factors that make securing users against phishing a challenging design problem and have derived design requirements for authentication schemes.

Studies that have gathered empirical data have drawn on **surveys** (3 publications), **lab** (20 publications), **online** (12 publications), or **field experiments** (16 publications) to analyze, e.g., the efficacy or usability of user-oriented phishing interventions. For instance, the effect of training material embedded in the process of sorting emails has been studied by Kumaraguru *et al.* [47], who have first employed a think-aloud vignette lab experiment, which has then been further tested in the field in the form of an online training game.

As for sample sizes, studies in our literature data range from small (< 20 participants) representative groups (e.g., [12, 36, 93]) to large-scale experiments with more than 1,000 participants (e.g., [47, 66, 85]). Field experiments were often conducted among university students and staff (e.g., [85]), rarely among non-university employees (e.g., [63]), or by evaluating real-world users' interactions with browser extensions or applications (e.g., [66]).

While most research articles in our sample have explored short-time effects of phishing interventions, some have employed longitudinal studies in order to investigate long-term effects. For example, Kumaraguru *et al.* [44] have observed knowledge retention of at least 28 days for users who had been trained via simulated phishing attacks and Silic & Lowry [73] have employed a long-term field experiment to investigate longitudinal effects of gamification on employees' intrinsic motivation to comply with security efforts.

With regard to the validity of experimental setups, previous works have pointed out that information security behavior research heavily relies on studying users' information security behavior as their primary activity on a computer [23,33,35]. In reality, however, responding to phishing threats is a secondary task that is embedded in a primary task, such as answering email or searching the internet. This leads to users facing the difficulty of switching between their primary and secondary activity, which may result in overlooking security warnings or disregarding educational offers. While many lab and online studies of our sample have studied their subjects' behavior as a primary task (e.g., by asking them to sort links into "legitimate" or "phishing" [5, 76]), others have assigned them fictional primary tasks to attend to. By using cover stories, such as sorting emails for a colleague or shopping online [43, 61], researchers have aimed to study phishing detection as a secondary task. However, it is arguable whether such artificial experimental setups can align with the complex nature of phishing. With regard to the realism of phishing experiments, Schechter *et al.* [68] have shown that role-playing participants behave less securely than those who act in a personal context (e.g., participants asked to log into a bank account with predefined passwords showed less secure behavior than those using their own passwords). While online or lab experiments are essential to test and refine theories of user behavior as well as to improve artifacts in human-computer interaction, conducting studies in a realistic environment is crucial to allow for robust and practice-oriented results. In our literature sample, less than one third (16 of 51) of experiments have been conducted in a real-world field setting.

3.2 A Taxonomy of User-Oriented Phishing Interventions

Our literature review has revealed that, while user-oriented phishing interventions all pursue one common goal (to protect users from phishing threats), they vary widely with regard

to their underlying concepts and intended effect. Prior literature has presented vague attempts of categorizations of phishing interventions. For example, Kirlappos & Sasse [43] have described two main approaches, namely anti-phishing indicators and user education, whereas Xiong *et al.* [94] have distinguished between warnings and training, and the integration of both. Similarly, Wash [85] has observed three styles of phishing interventions: general-purpose training messages that communicate "best practices", fake phishing campaigns, and in-the-moment warning messages. We chose to follow a fourth approach by Jansen & van Schaik [39], who have roughly described four different categories of user-oriented phishing interventions: **education**, **training**, **awareness-raising** and **design**. In their pure form, education and training interventions typically promote sustainable, long-term secure behavior, with the central aim that the application of knowledge and skills transfers to the real-world and enables users to engage in secure practices [79], whereas awareness-raising and design interventions aim to improve users' security during specific activities (such as logging into a website or reading an email) in the short term. Our literature analysis has revealed, however, that interventions often incorporate elements of more than one type.

Based on the literature data, we have derived a taxonomy of user-oriented phishing interventions as presented in Figure 2. In the following sections, we will describe the four categories and their respective subcategories in detail.

3.2.1 Education

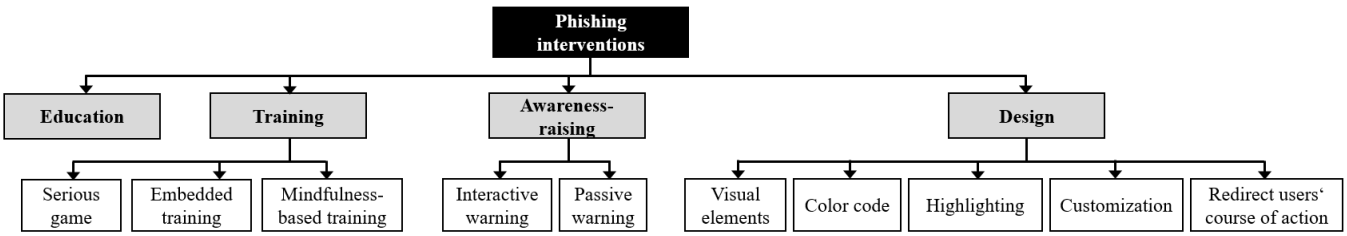
Purely educational interventions focus on developing knowledge and understanding of phishing threats and ways to mitigate them, e.g., by providing educational media, such as texts or videos, or by discussing online threats during in-class training. For this category, we have identified 7 publications in total. However, only three of them have considered education as a solitary intervention. For example, Wash & Cooper [85] have investigated which role the perceived origin of phishing education material plays in terms of effectiveness and have found that facts-and-advice-based training from perceived security experts surpasses the same training from peers. Four research works have studied phishing education in interaction with awareness-raising interventions by adding educational texts to fear appeals [39, 70] or warnings [95]. For example, Yang *et al.* [95] have found that a warning trigger combined with an educational text enhances its effectivity, whereas the educational element itself was not sufficient to provide phishing protection. Others have first provided extensive education in order to refer back to it during awareness-raising interventions later on [8]. Education interventions have been studied in rather traditional text-based, video-based or in-class formats. More progressive formats, such as online games, comprised interactive and hands-on exercises and were hence categorized as training.

3.2.2 Training

Compared with educational interventions, training goes one step further. It typically involves some kind of hands-on practice, where users develop skills that they can apply in case of a real threat. Since the term "training" is quite widespread in everyday language use, interventions that have been described as training by the respective authors might have been categorized as education within this work. Training approaches aim to enable users to identify phishing websites, phishing emails, or other malicious attacks. They employ interactive elements or exercises, where users can develop skills such as reading a URL, analyzing an email, or recognizing social engineering attempts. They often do so by exposing the user to a similar attack within a secure environment, either in an artificial or a real-world setup. Within our phishing literature data, about half of the publications (31 research papers) were dedicated to training interventions. Among them, we were able to distinguish several approaches.

Training interventions are typically rule-based. That is, their goal is to train individuals to identify certain cues to take protective action [76]. In our sample, 16 publications have explored such training in a **serious game** context, mostly taking place online and often focusing on teaching users how to identify phishing links by using cues in URLs (e.g., [5, 12, 72, 76]). For instance, Sheng *et al.* [72] have introduced "Anti-Phishing Phil", a game that is designed to teach users how to identify fraudulent websites based on the use of IP addresses, subdomains or deceptive domains in a URL. Similarly, "NoPhish" is a mobile app that guides users through several levels of analyzing and recognizing phishing URLs [12, 76]. The authors have found a long-term effect with regard to users' knowledge retention; that is, users who had played the NoPhish game have shown a better ability to decide upon the legitimacy of a URL. Silic & Lowry [73] have observed that gamified security training systems, which include elements such as levels or leader boards, enhance users' intrinsic motivation and yield better security behavior. Offline games have been explored in the form of board [6] or escape room [7] games.

Apart from gamified contexts, **embedded training** has gained momentum in recent phishing intervention research. Embedded training describes interventions that *"train a skill using the associated operational system including software and machines that people normally use"* [4, p. 406]. In other words, embedded training combines testing users' behavior in their normal personal or work environments with instant corrective performance feedback. It has been argued that the experience of "being phished" constitutes a so-called most teachable moment, where lasting change to attitudes and behaviors is possible [13]. Embedded training has been studied by 13 publications in our literature sample. As an example, "PhishGuru" is a program that simulates harmless but realistic phishing emails right into users' email inboxes [44–47]. When falling for a simulated phishing attempt (i.e., clicking



Category	Definition	Phishing interventions	Articles
Education	Educational interventions aim at developing knowledge and understanding of phishing and how to protect oneself against it.		
Education		Text-based, video-based, or in-class education	[8,39,48,63,70,85,95]
Training	Training interventions refer to interactive elements or exercises, which provide users with hands-on practice. They often take place by presenting a realistic phishing attempt within a secure environment.		
Serious game	Serious games refer to gamified contexts in which users can train how to recognize and analyze phishing attacks.	Online game (e.g., "NoPhish"), mobile app, board game, escape room game	[5–7,12,17,21,28,31,32,47,60,71–73,86,88]
Embedded training	Embedded training refers to training schemes that combine testing users' behavior in their normal environment with instant corrective performance feedback.	Phishing simulation in combination with a "teachable moment" (e.g., "PhishGuru")	[4,10,11,13,15,30,44–46,52,75,85,94]
Mindfulness-based training	Mindfulness-based approaches refer to trainings that increase users' awareness of context.	Approaches that teach users to dynamically allocate attention during message evaluation	[40]
Awareness-raising	Awareness-raising interventions refer to warnings that are placed in situ and raise users' awareness of potential phishing attempts during their primary course of action.		
Interactive warning	Interactive warnings refer to awareness-raising interventions that do require user interaction, i.e., interrupt the users' course of action.	Forced-attention warning, security questions, interactive fear appeal	[2,25,29,39,61,62,68,70,75,83,89,93,95,96,98]
Passive warning	Passive warnings refer to awareness-raising interventions that do not require user interaction.	Security toolbar, display of information on the legitimacy of a website	[8,25,92]
Design	Design interventions refer to design choices that aim at supporting or guiding users' behavior with respect to their secure handling of online activities.		
Visual elements	Visual elements refer to interventions that use the visual appearance of, e.g., a login form or website, to support users' security behavior.	UI dressing, dynamic security skins, trust logo, image	[24,34–36,43,49,51,68,81,97]
Color code	Color codes refer to simple visual cues for users to distinguish between secure and risky environments.	Traffic light colors	[43,89,92]
Highlighting	Highlighting refers interventions that draw users' attention towards critical elements.	Domain highlighting, sender highlighting, highlighting differences in out-of-focus tabs	[22,50,56,83]
Customization	Customization refers to interventions that let users customize the visual appearance of, e.g., a login form.	Custom icon, custom image, custom UI dressing	[24,34–36,51,68,81,97]
Redirect users' course of action	This category refers to interventions that redirect users' course of action, for example by offering more secure alternatives.	Browser sidebar for entering credentials, suggesting alternative websites, creating habit of using bookmarks, delayed password disclosure	[35,38,54,66,93]

Figure 2: A taxonomy of user-oriented phishing interventions.

on a phishing link), users were redirected to a training website explaining how phishing attacks work and how they can protect themselves from fraudulent emails and websites. Embedded training is a promising approach with regard to the real-world environment it takes place in: users are not in a training environment (such as an online game), but receive training only if they fall for a simulated phishing attempt during their everyday duties. Thus, knowledge and changes in security attitudes and behaviors can be transferred to real phishing attempts more easily. This is reflected in a growing business of embedded "phishing simulation training" by commercial information security companies [2]. Kumaraguru *et al.* have shown that training with "PhishGuru" helps users retain what they learned in the long term and that multiple training interventions increase performance [44].

Beyond rule-based training, Jensen *et al.* [40] have shown that expanding the rather conventional training toolkit with **mindfulness-based training** leads to a better ability to avoid phishing attacks. Mindfulness training teaches users to dynamically allocate attention during message evaluation (*"(1) Stop! (2) Think ... (3) Check."*) and aims to increase users' awareness of context. This method seems to be particularly effective for participants who were already confident in their detection ability.

3.2.3 Awareness-raising

The third category, awareness-raising, aims at focusing users' attention on potential threats and their countermeasures in situ,

[2]For example, Proofpoint ThreatSim® (proofpoint.com), Sophos Phish Threat (sophos.com), IT-Seal Awareness Academy (it-seal.de), Lucy Security (lucysecurity.com), and many others.

that is, as part of their primary course of action. Awareness-raising interventions might, for example, interrupt the user's workflow to set security-conscious behavior on their agenda. We have identified 17 studies of awareness-raising interventions, of which three explore **passive warnings** (i.e., the warning does not require user interaction), and 15 investigate on **interactive warnings** (i.e., the warning does require user interaction). Several prior studies have shown that passive interventions such as security toolbars in an internet browser are ineffective at preventing phishing attacks [25, 92].

Interactive warnings have been shown to have promising effects on users' phishing vulnerability. For example, the browser sidebar "Web Wallet" [93] acts as a secure way to submit sensitive information by suggesting alternative safe paths to intended websites and forcing users' attention by integrating security questions. Several research works have explored the mechanism of forced attention: Volkamer *et al.* [83] have introduced "TORPEDO", an email client add-in that delays link activation for a short period of time. As for web browser phishing warnings, Egelman *et al.* [25] have shown that interactive warnings, where users have to choose between options such as "Back to safety" or "Continue to Website", are heeded significantly more often compared to passive warnings. Furthermore, Petelka *et al.* [61] have shown that link-focused warnings are more effective than general email banner warnings in protecting users from clicking on malicious URLs, and that forced attention amplifies this effect. When comparing awareness-raising interventions that include educational elements (such as descriptions of the consequences of phishing, or explanations why a certain link or file is classified as potentially dangerous) to those that do not provide any additional information, the former were found to be more effective [75, 95]. Two research works have examined the potential of fear appeals, that is, short, informative messages that communicate threats, and have found that concrete fear appeals (compared with abstract fear appeals) are more effective to increase actual compliance behavior [39, 70]. This indicates that a combination of warning, forcing users' attention, and therein embedded tangible education yields a promising protection against phishing threats.

3.2.4 Design

Lastly, design choices can act as phishing interventions if they facilitate desirable user behavior [39]. We have identified 20 publications that investigate design interventions aimed at supporting users' secure handling of email and online activities.

Visual elements play a role in several research works (10 publications). For instance, the potential of "dynamic security skins" has been explored by Dhamija and Tygar [24], who have presented an authentication scheme where users rely on visual hashes from a trusted source that match the website background for legitimate websites.

Visual elements also come into play when offering users design options to **customize** security indicators, such as custom images or icons. An example is "Passpet", a browser extension by Yee & Sitaker [97] that acts as a password manager and an interactive custom indicator. Iacono *et al.* [36] have proposed so-called "UI-dressing", a mechanism that relies on the idea of individually dressed web applications (e.g., by using customized images) in order to support the user in detecting fake websites.

Color codes refer to simple visual cues (e.g., traffic light colors) for users to distinguish between secure and risky environments. They have, so far, been observed to be of limited success in the form of security indicators that signal whether a website is genuine or fake [43, 92]. Furthermore, Wiese *et al.* [89] have explored color codes in the context of email application UI design, where they were used to indicate the presence of digital signatures.

In contrast, **highlighting** draws users' attention to critical elements. For example, both Volkamer *et al.* [83] and Lin *et al.* [50] have investigated the effectiveness of domain highlighting in order to enable users to find the relevant part of a URL, whereas Nicholson *et al.* [56] have explored highlighting an email's sender name and address.

Other design interventions set out to **redirect users' course of action**, for example, by creating the habit of using browser bookmarks instead of hyperlinks to access sensitive websites such as login pages [35]. Ronda *et al.* [66] have developed "iTrustPage", a tool that warns the user about suspicious websites (e.g., a fake PayPal website). Beyond that, it offers corrective action in the form of suggesting alternative websites that are deemed trustworthy based on Google's search index (e.g., the real PayPal website).

Surprisingly, while the concept of digital nudging has gained widespread attention (among others in usable security research, e.g. [16, 19, 42, 100]) in recent years, only one article in our sample has investigated the effect of a nudge: Next to highlighting the name and address of an email's sender, Nicholson *et al.* [56] have investigated the effect of a social salience nudge ("62% of your colleagues received a version of this email") on users' phishing vulnerability. While several other design interventions contain nudge-like elements (such as color codes or highlighting), none of them have been designed as or labelled a nudge by the respective authors. We will further elaborate on the potential of digital nudging in phishing interventions in Section 4.2.

3.3 Which Phishing Attack Vector Does the Intervention Address?

While the term "phishing" originally describes cyber attacks that aim for users' passwords, it is now used to describe all sorts of attack vectors [23]. Those attack vectors differ in terms of the criminals' intended outcome (e.g., disclosure of confidential information or implanting malware) and the

user's primary action during which the attack takes place (e.g., clicking on a link or downloading a file). In the following, we will analyze the range of attack vectors that the phishing interventions in our sample aim to intervene in detail.

Phishers predominantly choose email messages as their first approach towards the user [85]. About 3.9 billion people worldwide have email accounts and collectively send and receive over 290 billion emails per day [37]. Email thus presents a means of communication that can easily be abused to take advantage of users' credulity by blending into daily personal or professional correspondence. Since attackers employ social engineering techniques (e.g., urgency cues or trustworthy-seeming visual elements) to elicit specific actions such as clicking a link, opening an attachment, or disclosing sensitive information, **deceptive email messages** themselves can be considered as an attack vector. Seventeen publications address users' ability to distinguish legitimate emails from phishing emails by paying attention to the email message itself. For instance, Caputo *et al.* [13] have studied embedded phishing training that aims at educating users on how to recognize phishing emails based on various criteria such as mismatched names, spelling mistakes, or intuition.

Phishing messages furthermore often offer a link, which, for example, might execute a drive-by download of ransomware [85] or redirect the user to a website masquerading as a legitimate login page. Previous research suggests that, after recipients click on a phishing link, they rarely detect subsequent fraudulent attempts such as a counterfeit login page or change their course of action [91]. **Disguised URLs** (such as, e.g., *paypa1.com*, *mybank.com-secure.biz*, or *tinyurl.com/XYZ*), that make the user believe that they are clicking on a reliable link, hence constitute a prominent attack vector. Accordingly, more than half of our literature sample (33 publications) explores user-oriented phishing interventions that aim at preventing users from clicking malicious links. These interventions mostly consider links in the context of an email. For example, Volkamer *et al.'s* [83] email client add-on "TORPEDO" uses tooltips to focus the user's attention on a link's domain. While links with whitelisted or previously visited domains will be activated immediately when clicked, "TORPEDO" will delay the activation of other links for a few seconds to encourage the user to check the URL's domain carefully. Several training games provide users with an in-depth explanation and exercise about how URLs can be obfuscated to mimic reputable sources, and have been shown to help users make better decisions concerning the legitimacy of URLs in the long term (e.g, [12, 72, 76]).

While links are usually accessed via clicking on a link, **QR codes** gain in popularity due to their ease of distribution and fast readability. Since the user has no means to examine the URL behind a QR code before scanning it, they constitute a hidden security threat. One single publication in our sample has addressed this issue by exploring security features of QR code scanners that help users to detect phishing attacks [96].

Besides disguised URLs, **imitated websites** can present another attack vector. For example, cyber criminals employ imitations of well-known websites in order to exploit users' trust in visually familiar or trustworthy environments. Ten publications in our literature sample have addressed this attack vector. For example, Iacono *et al.* [36] have proposed an intervention that relies on the idea that the whole appearance of a web application is dressable according to the user's individual preferences, raising users' attention for unofficial sites that do not align with the expected appearance. Regarding phishing interventions that are being displayed on websites, Kirlappos and Sasse [43] have revealed that arbitrary logos, certifications, or advertisements that do not imply trustworthiness of a website might have a higher reassurance to users than actual security indicators. This gives an example of how user-oriented interventions themselves can be exploited by cyber criminals to trick users into placing trust into a website.

When browsing the internet, interventions such as padlock icons or warning messages inform the user about a website's **SSL/TLS certificates**[3]. Interventions that inform or warn the user about SSL/TLS have been addressed, for example, by Reeder *et al.* [62] or Schechter *et al.* [68]. So-called man-in-the-middle attacks, where criminals use legitimate websites that do not encrypt data transmission by SSL/TLS to capture the user's sensitive data during an online transaction, have been a serious phishing attack vector in the past. Since nowadays, however, more than 80% of phishing sites have SSL/TLS encryption enabled [1], this attack vector will likely cease to play a role in the near future.

We now move from the preliminary stages (such as tricking users into trusting an email, link, or website) to the centerpiece of a phishing attack. One central aspiration of cyber criminals is to lure their victim into disclosing sensitive information, e.g., login credentials. Accordingly, several prior works (12 in our literature sample) have studied interventions that address the process of users' **authentication**. For example, Dhamija & Tygar [24] have introduced an interaction technique for authentication that provides a trusted window in the browser dedicated to username and password entry, which uses a photographic image to create a trusted path between the user and password entry fields. Similarly, Yee & Sitaker's [97] browser extension "Passpet" constitutes a password manager that helps users securely identify trustworthy login forms.

Besides fishing for credential data, phishers' efforts are directed at prompting the user to download or execute **malware**, that is, malicious software that can harm the user's device or their entire network. Malware attacks have rapidly grown over the recent years, e.g., in the form of ransomware attacks [74]. Surprisingly, interventions that aim at preventing users from executing malware are scarce in our literature data. Only three publications have addressed this attack vector: Wen *et al.* [88]

[3] Transport Layer Security (TLS), and its predecessor, Secure Sockets Layer (SSL), are cryptographic protocols designed to provide communications security over a computer network [57]

have included different kinds of potentially malicious attachments in their conception of a role-play anti-phishing training game, whereas Reeder *et al.* [62] have explored users' interaction with browser warnings that warn against downloading malware. Reinheimer *et al.* [63] have taught how to identify dangerous files in their in-class training.

Malicious **mobile applications** can act as a phishing attack vector, for example, by masquerading as a legitimate online banking app. One publication in our sample has discussed personalized security indicators in mobile applications [51].

In addition to the above-described investigations of specific phishing attack vectors, 10 publications have approached the topic of phishing in a more general manner. Most of these publications have examined training formats, such as online games, that cover the phenomenon of **phishing in a broader sense** without addressing or intervening one attack vector in particular.

Figure 3 illustrates the distribution of our literature data across different phishing attack vectors. Since some publications address interventions to more than one phishing attack vector, the sum of the displayed data points is larger than the literature sample size of 64 articles. For a detailed categorization of all articles, please refer to Table 2 in the appendix.

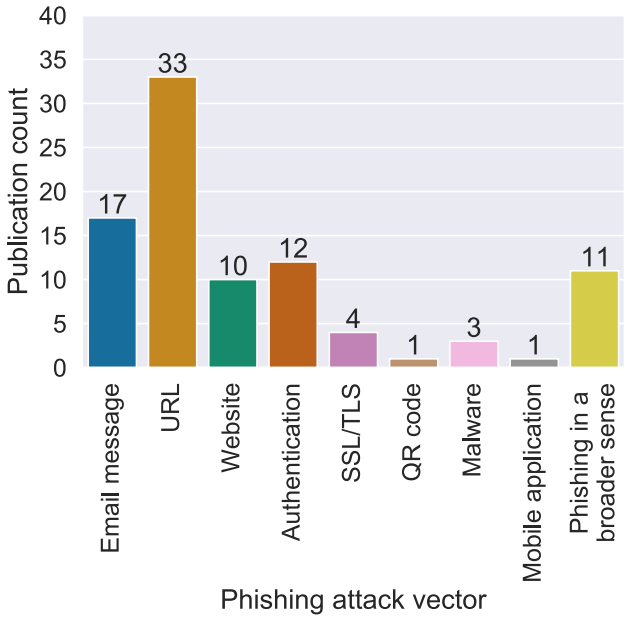


Figure 3: Overview of the attack vectors addressed by phishing interventions studied in the literature data.

3.4 When Does the Intervention Take Place?

Diving deeper into the analysis of user-oriented phishing interventions, we have further considered the point in time at which the intervention takes place. We have found that many interventions, mostly those aiming at training or education, are designed to take place as a precautionary measure, often long before the user interacts with a potential phishing context. We have identified 23 articles that present such interventions and have labeled them as **pre-decision interventions**. For instance, Jansen & van Schaik [39] have shown that confronting users with fear appeal messages is suitable to heighten their cognitions, attitudes, and intentions with regard to secure online behavior. Furthermore, all kinds of non-embedded education or training (e.g., in-class education [48], online games [72], mobile training apps [12]) clearly take place pre-decision.

Most of the approaches in our literature sample focus on interventions that take place during users' course of action, that is, **during the user's decision** between phishing and legitimate content in a real-world context. Those 31 articles mostly describe awareness-raising and design interventions, sometimes combined with educational elements. For instance, Petelka *et al.* [61] have examined the effectiveness of different levels of link-focused warnings when sorting emails, whereas various design interventions such as color codes, customization or highlighting aim to support users' decisions during their course of action.

We have further identified 11 publications describing interventions that take place **post-decision**, that is, after a user's decision on potential phishing contents was already made. This goes especially for embedded training, where training follows right after the user has been "phished" by a simulated attack.

Combinations of pre-, post-, and during decision intervention have been studied only once in our sample: Blythe *et al.* [8] have introduced an approach that consists of initial video-based education, which is then referred back to by security warnings during the users' individual course of action.

While several research works have employed longitudinal studies to examine the long-term effects of user-oriented phishing interventions (see Section 2), little has been investigated on interventions that take place regularly, e.g., by giving regular warnings or recurringly providing users with training. Reinheimer *et al.* [63] have explored the effect of reminding users of initial phishing awareness education and have found that reminders after half a year are recommended and that measures based on videos or interactive examples perform better than text-based reminders. Furthermore, several embedded training interventions have been explored in terms of the effect of recurringly simulated phishing emails (e.g., [13,15,44,52]).

Figure 4 sums up the distribution of the time of intervention across our literature sample.

3.5 Does the Intervention Require User Interaction?

Beyond the categorization as presented in Figure 2, we have analyzed all interventions in terms of whether they require active user interaction, e.g., whether the user's workflow is

Figure 4: The time at which the intervention takes place in relation to the user's decision, across our literature sample.

interrupted by the intervention and whether the user can only proceed when undertaking a certain action or decision. These interventions were classified as **interactive**. In contrast, interventions that only provide information or feedback to the user without actively interrupting their workflow are deemed **passive** interventions. Some of the 64 articles in our literature sample have addressed both interactive and passive interventions.

Across our sample, 48 publications describe phishing interventions that require user interaction. We mainly divide between two kinds of interactive interventions, one being interactive warnings as described in Section 3.2.3, which usually require a few seconds of the user's time and attention before they can proceed with the task at hand (e.g., [25, 61]). The other subset is formed by training and education approaches (see Sections 3.2.1, 3.2.2), which commonly require the user to actively engage in an exercise for at least several minutes up to hours, for example, online training games [12, 31, 72] or in-class training [48]. A total of 16 interventions can be described as passive, including passive warnings (e.g., [92]), some educational interventions (e.g., [39]), and also several interventions belonging to the design category. As an example, we have classified domain highlighting [50] as passive, since it does not require any interaction on the user's side and can also be easily ignored, or even overlooked, by the user.

4 Discussion

In the previous section, we have examined a plethora of user-oriented phishing interventions from various angles and have revealed surprising and relevant insights. Above all, we have found a highly fragmented landscape of educational interventions, training, awareness-raising warnings, and anti-phishing designs, which users need to navigate through when being pushed towards secure online behavior. To summarize and connect the findings across the dimensions of analysis, Figure 5 displays an integrative plot of all phishing interventions in our sample. Getting back to our research questions *RQ1* and *RQ2*, we devote the remainder of this article to discussing our findings and positioning them in current usable security research. After looking at the user effort and intrusiveness of prior phishing interventions in Section 4.1, we discuss the potential of digital nudges regarding phishing prevention in Section 4.2. We then address the role of users' cognitive processes when dealing with potential security threats in Section 4.3. Further, we consider the imbalance of phishing attack vectors addressed by prior intervention research in Section 3.3, and discuss the potential of tailored phishing interventions in Section 4.5. Subsequently, we highlight methodological aspects in Section 4.6, and lastly address limitations of our work in Section 4.7. We then sum up our contributions in Section 5, including an overview of our propositions for future phishing intervention research.

4.1 User Effort and Intervention Intrusiveness

One particularly salient finding is that most user-oriented phishing interventions encumber the user with additional effort with respect to their workload and time, for example, in the form of playing a training game [72], interacting with embedded training [44], answering security questions [93], or waiting for delayed link activation [83]. Those seconds or minutes required to interact with an intervention cumulatively drain time from individual and organizational productivity. Moreover, they often intrusively disrupt the user in their primary goals, hence again substantially decreasing productivity by distraction and potentially leading to stress and frustration. This aligns with Sasse's [67] observation that user time and effort are rarely at the forefront of security studies and that the issue of user effort and intrusiveness has scarcely been considered. Sasse has argued that designers of security tasks should focus on *"causing minimum friction"* and *"must acknowledge and support human capabilities and limitations"* [67, p. 82]. She has called for subjecting security measures to a cost-benefit test and to give up on perfection and focus on essentials. On the other hand, passive, that is, less intrusive interventions have been observed to be of limited success as of yet [36, 43, 68, 92]. It hence remains the most challenging task to design effective user-oriented phishing interventions that prove themselves usable in individuals' everyday online activities, particularly with regard to user effort and intrusiveness. Digital nudging [77, 87] might constitute an unintrusive yet promising approach for this endeavor. In Section 4.2, we evaluate which elements of prior, effective interventions could be classified as nudges retrospectively and present ideas for future approaches. As for training and education interventions, Cranor & Garfinkel [20] have argued that *"the world's future cyber-security depends upon the deployment of security technology that can be broadly used by untrained computer users"*, hence questioning the usability of such approaches. It is still an open question whether interventions need to be understood by the user (e.g., via providing educational information) in order to be effective [25, 100], whereas it has been observed that intervention clearness (e.g., with regard to their message concreteness [70] or their location [61]) increases effectiveness. This spans an interesting research area with potentially crucial insights for the design of future phishing interventions.

Figure 5: Overview of user-oriented phishing intervention literature, spanned by attack vector, time of intervention, and intervention category. Since some articles have addressed several attack vectors or intervention categories, they appear more than once.

4.2 Digital Nudges As Phishing Interventions?

As described in Section 3.2.4, the concept of digital nudging has scarcely been drawn on in phishing intervention research as of yet. The term nudging has been introduced by Thaler & Sunstein [77] in 2009. Digital nudges describe user-interface design elements that target automatic cognitive processes, such as biases or heuristics, to gently push end-users, with little mental effort, to perform the "right" behavior without limiting their choice set [77, 87]. In this section, we aim to discuss the potential of digital nudges in phishing intervention research, especially since prior research in related fields, such as digital privacy-protection or security choices [16, 65, 100], can serve as a solid basis to start from. Surprisingly, phishing intervention literature from 2018 onward has focused on education, training, and awareness-raising measures, while neglecting design interventions (see Table 2 in the appendix). Design phishing interventions might provide significant value to users' security if they succeed in nudging users towards secure behavior, while not being perceived as intrusive with regard to their primary goals.

In an extensive review, Caraban *et al.* [14] have classified six distinct nudge categories in the area of human-computer interactions. In the following, we exemplarily discuss how existing interventions make use of several of those mechanisms already (although not labeling them as nudging) and present novel ideas on how digital nudging could be applied in future phishing intervention research.

Facilitate. Facilitating nudges use mechanisms to lessen users' effort. In our sample, highlighting domains [50, 83] or sender addresses [25] falls in this category since it makes it easier for users to spot the relevant part of an URL or email sender. We propose to take this approach further, for example, by displaying a link's domain next to the link text in an email, with only the domain being clickable.

Confront. Confronting nudges aim to create friction by throttling users' mindless activity or reminding them of the consequences. Several of the interventions in our sample can be described as such, for example, interactive awareness-raising measures as described in Section 3.2.3. As we have argued in the previous section, burdening the user with intrusive distraction and effort cannot be an efficient answer to current and future challenges in cyber security. We hence argue that confronting nudges should be designed to be of minimal possible friction. For example, they could remind the user of consequences by making security risks tangible.

Deceive. Deceptive nudges influence the perception of the available options, e.g., by adding inferior alternatives or placebos. None of the analyzed interventions could be sorted into this category, and we do not deem deceptive nudges suitable for phishing intervention research.

Social Influence. This type of nudge makes use of social influences on people's choices. Examples of social influence within the analyzed articles include the comparison of facts and stories provided by peers vs. experts on anti-phishing education [52] as well as Nicholson *et al.'s* [56] social saliency nudge. Furthermore, social influence has been studied in a social learning environment in terms of gamified elements such as levels or leader boards [73]. Future social influence nudges could provide users with information on, e.g., their vs. their peers' performance in phishing simulations or incident reporting activities.

Fear. Two research works of our sample [39, 70] have introduced fear appeals as phishing interventions with promising results regarding users' protection motivation, attitudes, intentions and compliant behavior. However, both articles have studied fear appeals far from a real-world scenario, using text-based treatments and a survey instrument. We suggest that fear nudges, which, integrated in the user's course of action, aim to invoke fear to encourage a certain choice, are of high interest for future research. Nevertheless, they require ethical considerations [64]. As an example, we imagine a brief but concrete [70] and strong [39] fear appeal next to email

attachments, addressing the risk in terms of financial losses and operational damage coming along with this file type and a potential malware infection. The fear appeal could be framed positively to address ethical concerns by showing how the user could protect against these threats easily.

Reinforce. Reinforcing nudges aim to support certain behaviors, e.g., by ambient feedback or just-in-time prompts. Regarding the first, we found mechanisms ranging from color-coding security indicators on websites [34, 92] to providing customized background images [51, 68] in our sample. One shortcoming of these interventions seems to be that users cannot distinguish between legitimate security indicators (such as a color code) and untrustworthy signs, such as arbitrary logos and certifications [43]. One way to battle this could be to make ambient feedback more comprehensive or standardized, e.g., by color-coding complete email or website windows. Concerning just-in-time prompts, in order to condense prior warning interventions to the pure form of a digital reinforcement nudge, we ideate an authentication intervention that displays the domain of a login website above any login form when placing the cursor in the login field.

Finally, suitable nudges could be easily combined with other interventions types, for example, educational elements [100], as shown by successful examples [39, 70, 95]. As illustrated in Table 2 in the appendix, interventions that combine educational with awareness-raising or design approaches have rarely been studied in phishing research as of yet.

4.3 Shifting Users' Cognitive Frame

From a different perspective, Wash [84] has adduced IT experts' approach towards identifying phishing emails and has observed that experts naturally follow a three-stage process: (1) making sense of the email, relating it to one's personal context, and deriving required action (2) becoming suspicious and investigating, and (3) dealing with the email by deleting or reporting it. He argues that shifting the user's cognitive frame from sensemaking to investigation is crucial for the success of phishing prevention measures. However, half of the interventions in our literature sample have addressed training or education measures (see Figure 5). Those mostly neglect the initial process of noticing slight discrepancies or cues in an email in the sensemaking frame and provide support only in the investigation frame (e.g., how to analyze an URL). While Jensen *et al.'s* [40] mindfulness-based training aims to support users in their awareness of context, and such during their sensemaking process, long-term efficacy is uncertain.

At the same time, users' own security goals should not be neglected: Kirlappos *et al.* [43] have argued that users do not focus on security warnings, but rather look for signs to confirm a website's trustworthiness. For example, users have been shown to trust websites that display advertisements affiliated with known entities or those with familiar website layouts - while both factors do not give evidence of the website's trustworthiness. Therefore, the authors have called for security education to consider the drivers of users' behavior in their respective situation and, conversely, to eliminate users' misconceptions that lead to insecure behavior.

We hence argue that future phishing interventions should strive to meet the user in their own respective sensemaking process, for example, when reading emails, shopping, or doing bank transactions online. Digital nudges might play an important role in this particular case, as well. Supporting the user's cognitive frameshift from the stage of sensemaking to the stage of investigating if certain cues or discrepancies are present will be an important path for future research and will complement the diverse landscape of education and training measures.

4.4 What About Malware?

Regarding the phishing attack vectors addressed across our literature data, we have found that more than half of the interventions focus on the attack vector URL, for example, by training users' skills in analyzing a link or raising their awareness in situ. Interventions supporting the user with deceptive email messages, disguised websites, and fraudulent authentication forms follow by far (see Figure 5).

Malware poses a tremendous risk through current cyber attack patterns [18, 58]. Those attacks are often delivered by archive files or Microsoft Office documents which mimic, e.g., legitimate invoices. Since the user needs to download and open these files on their system, this presents quite a different attack procedure compared with clicking a link. Therefore, it is striking that only three publications have included educational, training, or awareness-raising interventions in their works that address malware alongside other attack vectors. None of the articles in our sample has focused on studying interventions that primarily support users in detecting or handling malware, nor have the challenges of malware interventions compared to previous phishing intervention research been addressed.

We therefore strongly suggest further research to expand previous approaches on phishing interventions in terms of the attack vector by taking into account malicious files and developing interventions that address the actual threat landscape.

4.5 Tailored Interventions

In the context of user interventions in cyber security, several studies have pointed out the potential of personalization regarding user traits [26, 41, 59], or the importance of context (e.g., personal vs. organizational [70]). It has been argued that using tailored instead of one-size-fits-all interventions may enhance their efficacy and user compliance [26].

Interestingly, our literature review does not reveal a strong focus on tailored user interventions to prevent phishing attacks. However, some of the approaches were indeed imple-

mented for specific target groups – mainly for rather heterogeneous groups of employees [73], or children [48]. Since spear phishing attacks are specifically targeted at personal or contextual vulnerabilities, considering users' traits, capabilities and requirements when developing and evaluating user interventions may be a decisive factor for their efficacy, suggesting a scope for future research.

4.6 Methodological Aspects

As described in Section 3.1, current research often lacks realism regarding the experimental setup since it remains challenging to study a phenomenon of deception that usually takes place during users' secondary tasks. Therefore, we argue that future research should not only focus on designing user-oriented phishing interventions, but also on developing experimental setups that account for a realistic analysis of users' security behavior.

Furthermore, we have found that the effect of recurring interventions has been studied scarcely (see Section 3.4). However, many interventions in our sample are designed to train, warn or guide users recurringly. Factors such as habituation [80] or security fatigue hence could have important effects. This proves another major shortcoming in prior phishing intervention research, which should be considered by future works.

4.7 Limitations

In this work, we have carefully selected (usable) security-specific databases to include a large number and variety of publications. Furthermore, the chosen search term was rather broad, and additional sources (such as security conferences) were considered to avoid overlooking relevant findings. Nevertheless, the list of publications analyzed in this research is probably not exhaustive. Furthermore, the features of the different phishing interventions were described in varying detail due to the individual focus and comprehensiveness of the articles. It is thus possible that certain interventions were classified differently by us than the authors themselves would have classified them. Therefore, this systematization of knowledge does not serve as an endpoint but as a starting point for identifying the current state, potential research gaps, and relevant paths for future work. We hope to not only provide a relevant summary and systematization of existing strategies for usable security-related researchers and practitioners but especially to encourage future studies in this increasingly relevant domain, where the human factor plays an essential role.

5 Conclusion

Phishing does not cease to be a threat to both personal and organizational data and operational security. It directly targets the human factor via deceptive emails, attachments, and websites, hence calling for user-oriented interventions that support individuals in recognizing and fending off such attacks. In this work, we have systematically analyzed 64 phishing intervention research articles for methodology, intervention type, attack vector, intervention time and user interaction, and have derived a taxonomy of user-oriented phishing interventions. Connecting the findings across the dimensions of analysis, as well as taking into account current movements in usable security research, we have revealed relevant insights and potential avenues for future work. The latter can be summarized as follows:

Minimize user effort and intervention intrusiveness. How can we design effective phishing interventions that cause minimum friction with the user's course of action and do not cumulatively burden the user with secondary time and workload? Which role does educational information play in intervention effectiveness, compared with intervention clearness and concreteness?

Explore the potential of digital nudging. How can facilitating, confronting, reinforcing, fear, or social influence nudges support users' course of action with regard to secure online behavior?

Help users shift their cognitive frame. How can we support users in the cognitive process of shifting from their primary goal of sensemaking towards noticing discrepancies if "something is off"? How can we transfer experts' expertise with phishing detection into effective end-user interventions?

Protect users from malware attacks. Which kinds of interventions can help to protect users from malware attacks? Which novel challenges do arise for malware-focused interventions, compared with threats employing malicious URLs or websites?

Explore tailored interventions. How can tailored phishing interventions enhance previous approaches?

Develop realistic experimental setups and study long-term effects. Which novel ways can be employed to align experimental setups with the nature of phishing and to account for longitudinal effects?

With this article, we hope to provide a comprehensive starting point as well as inspiration for future user-oriented phishing intervention research.

6 Acknowledgements

This research work has been funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] Greg Aaron. APWG Phishing Activity Trends 3rd Quarter Report 2020, 2020.

[2] Ahmed Abbasi, F Mariam Zahedi, and Yan Chen. Phishing susceptibility: The good, the bad, and the ugly. In 2016 IEEE Conference on Intelligence and Security Informatics (ISI), pages 169–174. IEEE, 2016.

[3] Luca Allodi, Tzouliano Chotza, Ekaterina Panina, and Nicola Zannone. The need for new antiphishing measures against spear-phishing attacks. IEEE Security & Privacy, 18(2):23–34, 2019.

[4] Abdullah Alnajim and Malcolm Munro. An anti-phishing approach that uses training intervention for phishing websites detection. In 2009 Sixth International Conference on Information Technology: New Generations, pages 405–410. IEEE, 2009.

[5] Nalin Asanka Gamagedara Arachchilage, Steve Love, and Konstantin Beznosov. Phishing threat avoidance behaviour: An empirical investigation. Computers in Human Behavior, 60:185–197, 2016.

[6] Malak Baslyman and Sonia Chiasson. "Smells phishy?": An educational game about online phishing scams. In 2016 APWG Symposium on Electronic Crime Research (eCrime), pages 1–11. IEEE, 2016.

[7] Erwan Beguin, Solal Besnard, Adrien Cros, Barbara Joannes, Ombeline Leclerc-Istria, Alexa Noel, Nicolas Roels, Faical Taleb, Jean Thongphan, Eric Alata, et al. Computer-security-oriented escape room. IEEE Security & Privacy, 17(4):78–83, 2019.

[8] Jim Blythe, Jean Camp, and Vaibhav Garg. Targeted risk communication for computer security. In Proceedings of the 16th international conference on Intelligent user interfaces, pages 295–298, 2011.

[9] Jan vom Brocke, Alexander Simons, Bjoern Niehaves, Bjorn Niehaves, Kai Reimer, Ralf Plattfaut, and Anne Cleven. Reconstructing the giant: On the importance of rigour in documenting the literature search process. In Proceedings of the European Conference on Information Systems (ECIS) 2009, pages 1–12, 2009.

[10] AJ Burns, M Eric Johnson, and Deanna D Caputo. Spear phishing in a barrel: Insights from a targeted phishing campaign. Journal of Organizational Computing and Electronic Commerce, 29(1):24–39, 2019.

[11] Mary B Burns, Alexandra Durcikova, and Jeffrey L Jenkins. What kind of interventions can help users from falling for phishing attempts: A research proposal for examining stage-appropriate interventions. In 2013 46th Hawaii International Conference on System Sciences, pages 4023–4032. IEEE, 2013.

[12] Gamze Canova, Melanie Volkamer, Clemens Bergmann, and Benjamin Reinheimer. NoPhish app evaluation: lab and retention study. In NDSS workshop on usable security, 2015.

[13] Deanna D Caputo, Shari Lawrence Pfleeger, Jesse D Freeman, and M Eric Johnson. Going spear phishing: Exploring embedded training and awareness. IEEE Security & Privacy, 12(1):28–38, 2013.

[14] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–15, 2019.

[15] Anthony Carella, Murat Kotsoev, and Traian Marius Truta. Impact of security awareness training on phishing click-through rates. In 2017 IEEE International Conference on Big Data (Big Data), pages 4458–4466. IEEE, 2017.

[16] Eun Kyoung Choe, Jaeyeon Jung, Bongshin Lee, and Kristie Fisher. Nudging people away from privacy-invasive mobile apps through visual framing. In Proceedings of the IFIP Conference on Human-Computer Interaction, pages 74–91, Berlin/Heidelberg, Germany, 2013. Springer.

[17] Gokul CJ, Sankalp Pandit, Sukanya Vaddepalli, Harshal Tupsamudre, Vijayanand Banahatti, and Sachin Lodha. Phishy - A serious game to train enterprise users on phishing awareness. In Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts, pages 169–181, 2018.

[18] Cofense. COFENSE Q3 Phishing Review. https://go.cofense.com/wp-content/uploads/pdf/Cofense-Q3_2020_Phishing-Review-report.pdf, 2020. Accessed: 2021-02-15.

[19] Lynne Coventry, Pam Briggs, Debora Jeske, and Aad van Moorsel. SCENE: A structured means for creating and evaluating behavioral nudges in a cyber security environment. In International conference of design, user experience, and usability, pages 229–239. Springer, 2014.

[20] Lorrie Faith Cranor and Simson Garfinkel. Security and usability: designing secure systems that people can use. O'Reilly Media, Inc., 2005.

[21] Tom Cuchta, Brian Blackwood, Thomas R Devine, Robert J Niichel, Kristina M Daniels, Caleb H Lutjens, Sydney Maibach, and Ryan J Stephenson. Human Risk Factors in Cybersecurity. In Proceedings of the 20th Annual SIG Conference on Information Technology Education, pages 87–92, 2019.

[22] Philippe De Ryck, Nick Nikiforakis, Lieven Desmet, and Wouter Joosen. Tabshots: Client-side detection of tabnabbing attacks. In Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security, pages 447–456, 2013.

[23] Alan R Dennis and Randall K Minas. Security on autopilot: Why current security theories hijack our thinking and lead us astray. ACM SIGMIS Database: the DATABASE for Advances in Information Systems, 49(SI):15–38, 2018.

[24] Rachna Dhamija and J Doug Tygar. The battle against phishing: Dynamic security skins. In Proceedings of the 2005 symposium on Usable privacy and security, pages 77–88, 2005.

[25] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 1065–1074, 2008.

[26] Serge Egelman and Eyal Peer. The myth of the average user: Improving privacy and security systems through individualization. Proceedings of the 2015 New Security Paradigms Workshop, pages 16–28, 2015.

[27] EUROPOL. World's most dangerous malware EMOTET disrupted through global action. https://www.europol.europa.eu/newsroom/news/worlds-most-dangerous-malware-emotet-disrupted-through-global-action. Accessed: 2021-02-08.

[28] Rubia Fatima, Affan Yasin, Lin Liu, and Jianmin Wang. How persuasive is a phishing email? A phishing game for phishing awareness. Journal of Computer Security, 27(6):581–612, 2019.

[29] Sophie Gastellier-Prevost, Gustavo Gonzalez Granadillo, and Maryline Laurent. A dual approach to detect pharming attacks at the client-side. In 2011 4th IFIP International Conference on New Technologies, Mobility and Security, pages 1–5. IEEE, 2011.

[30] Kristen K Greene, Michelle P Steves, Mary F Theofanos, and Jennifer Kostick. User context: an explanatory variable in phishing susceptibility. In in Proc. 2018 Workshop Usable Security, 2018.

[31] M Hale and R Gamble. Toward increasing awareness of suspicious content through game play. In 2014 IEEE World Congress on Services, pages 113–120. IEEE, 2014.

[32] Matthew L Hale, Rose F Gamble, and Philip Gamble. CyberPhishing: A game-based platform for phishing awareness testing. In 2015 48th Hawaii International Conference on System Sciences, pages 5260–5269. IEEE, 2015.

[33] Farkhondeh Hassandoust, Angsana A Techatassanasoontorn, and Harminder Singh. Information Security Behaviour: A Critical Review and Research Directions. In Proceedings of the European Conference on Information Systems (ECIS) 2020, 2020.

[34] Amir Herzberg and Ahmad Jbara. Security and identification indicators for browsers against spoofing and phishing attacks. ACM Transactions on Internet Technology (TOIT), 8(4):1–36, 2008.

[35] Amir Herzberg and Ronen Margulies. Forcing Johnny to login safely. Journal of Computer Security, 21(3):393–424, 2013.

[36] Luigi Lo Iacono, Hoai Viet Nguyen, Tobias Hirsch, Maurice Baiers, and Sebastian Möller. UI-Dressing to detect Phishing. In 2014 IEEE Intl Conf on High Performance Computing and Communications, 2014 IEEE 6th Intl Symp on Cyberspace Safety and Security, 2014 IEEE 11th Intl Conf on Embedded Software and Syst (HPCC, CSS, ICESS), pages 747–754. IEEE, 2014.

[37] The Radicati Group Inc. Email statistics report, 2019-2023 Executive Summary. https://www.radicati.com/?download=email-statistics-report-2019-2023, 2019. Accessed: 2021-02-08.

[38] Markus Jakobsson and Steven Myers. Delayed password disclosure. ACM SIGACT News, 38(3):56–75, 2007.

[39] Jurjen Jansen and Paul van Schaik. The design and evaluation of a theory-based intervention to promote security behaviour against phishing. International Journal of Human-Computer Studies, 123:40–55, 2019.

[40] Matthew L Jensen, Michael Dinger, Ryan T Wright, and Jason Bennett Thatcher. Training to mitigate phishing attacks using mindfulness techniques. Journal of Management Information Systems, 34(2):597–626, 2017.

[41] Debora Jeske, Lynne Coventry, and Pam Briggs. Nudging whom how : IT proficiency , impulse control and secure behaviour. In Proceedings of the CHI Workshop on Personalizing Behavior Change Technologies, pages 1–4, 2014.

[42] Shipi Kankane, Carlina DiRusso, and Christen Buckley. Can we nudge users toward better password management? an initial study. In Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems, pages 1–6, 2018.

[43] Iacovos Kirlappos and M Angela Sasse. Security education against phishing: A modest proposal for a major rethink. IEEE Security & Privacy, 10(2):24–32, 2011.

[44] Ponnurangam Kumaraguru, Justin Cranshaw, Alessandro Acquisti, Lorrie Cranor, Jason Hong, Mary Ann Blair, and Theodore Pham. School of phish: a real-world evaluation of anti-phishing training. In Proceedings of the 5th Symposium on Usable Privacy and Security, pages 1–12, 2009.

[45] Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Protecting people from phishing: the design and evaluation of an embedded training email system. In Proceedings of the SIGCHI conference on Human factors in computing systems, pages 905–914, 2007.

[46] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Lessons from a real world evaluation of anti-phishing training. In 2008 eCrime Researchers Summit, pages 1–12. IEEE, 2008.

[47] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Teaching Johnny not to fall for phish. ACM Transactions on Internet Technology (TOIT), 10(2):1–31, 2010.

[48] Elmer Lastdrager, Inés Carvajal Gallardo, Pieter Hartel, and Marianne Junger. How Effective is Anti-Phishing Training for Children? In Proceedings of the Thirteenth USENIX Conference on Usable Privacy and Security, SOUPS '17, page 229–239, USA, 2017. USENIX Association.

[49] Linfeng Li, Marko Helenius, and Eleni Berki. A usability test of whitelist and blacklist-based anti-phishing application. In Proceeding of the 16th International Academic MindTrek Conference, pages 195–202, 2012.

[50] Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, and John Aycock. Does domain highlighting help people identify phishing sites? In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pages 2075–2084, 2011.

[51] Claudio Marforio, Ramya Jayaram Masti, Claudio Soriente, Kari Kostiainen, and Srdjan Čapkun. Evaluation of personalized security indicators as an anti-phishing mechanism for smartphone applications. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pages 540–551, 2016.

[52] John Marsden, Zachary Albrecht, Paula Berggren, Jessica Halbert, Kyle Lemons, Anthony Moncivais, and Matthew Thompson. Facts and Stories in Phishing Training: A Replication and Extension. In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–6, 2020.

[53] Matthew DF McInnes, David Moher, Brett D Thombs, Trevor A McGrath, Patrick M Bossuyt, Tammy Clifford, Jérémie F Cohen, Jonathan J Deeks, Constantine Gatsonis, Lotty Hooft, et al. Preferred reporting items for a systematic review and meta-analysis of diagnostic test accuracy studies: the PRISMA-DTA statement. Jama, 319(4):388–396, 2018.

[54] Daisuke Miyamoto, Takuji Iimura, Gregory Blanc, Hajime Tazaki, and Youki Kadobayashi. Eyebit: Eye-tracking approach for enforcing phishing prevention habits. In 2014 Third International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS), pages 56–65. IEEE, 2014.

[55] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, Prisma Group, et al. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS med, 6(7):e1000097, 2009.

[56] James Nicholson, Lynne Coventry, and Pam Briggs. Can We Fight Social Engineering Attacks by Social Means? Assessing Social Salience as a Means to Improve Phish Detection. In Proceedings of the Thirteenth USENIX Conference on Usable Privacy and Security, SOUPS '17, page 285–298, USA, 2017. USENIX Association.

[57] Rolf Oppliger. SSL and TLS: Theory and Practice. Artech House, 2016.

[58] Constantinos Patsakis and Anargyros Chrysanthou. Analysing the fall 2020 Emotet campaign. arXiv preprint arXiv:2011.06479, 2020.

[59] Eyal Peer, Serge Egelman, Marian Harbach, Nathan Malkin, Arunesh Mathur, and Alisa Frik. Nudge

Me Right: Personalizing Online Nudges to People's Decision-Making Styles. SSRN Electronic Journal, 2019.

[60] Evan K Perrault. Using an interactive online quiz to recalibrate college students' attitudes and behavioral intentions about phishing. Journal of Educational Computing Research, 55(8):1154–1167, 2018.

[61] Justin Petelka, Yixin Zou, and Florian Schaub. Put your warning where your link is: Improving and evaluating email phishing warnings. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–15, 2019.

[62] Robert W Reeder, Adrienne Porter Felt, Sunny Consolvo, Nathan Malkin, Christopher Thompson, and Serge Egelman. An experience sampling study of user reactions to browser warnings in the field. In Proceedings of the 2018 CHI conference on human factors in computing systems, pages 1–13, 2018.

[63] Benjamin Reinheimer, Lukas Aldag, Peter Mayer, Mattia Mossano, Reyhan Duezguen, Bettina Lofthouse, Tatiana von Landesberger, and Melanie Volkamer. An investigation of phishing awareness and education over time: When and how to best remind users. In Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020), pages 259–284, 2020.

[64] Karen Renaud and Marc Dupuis. Cyber security fear appeals: Unexpectedly complicated. In Proceedings of the New Security Paradigms Workshop, pages 42–56, 2019.

[65] Karen Renaud and Verena Zimmermann. Nudging folks towards stronger password choices: providing certainty is the key. Behavioural Public Policy, 3(2):228–258, 2019.

[66] Troy Ronda, Stefan Saroiu, and Alec Wolman. Itrustpage: a user-assisted anti-phishing tool. ACM SIGOPS Operating Systems Review, 42(4):261–272, 2008.

[67] Angela Sasse. Scaring and bullying people into security won't work. IEEE Security & Privacy, 13(3):80–83, 2015.

[68] Stuart E Schechter, Rachna Dhamija, Andy Ozment, and Ian Fischer. The emperor's new security indicators. In 2007 IEEE Symposium on Security and Privacy (SP'07), pages 51–65. IEEE, 2007.

[69] Guido Schryen, Gerit Wagner, Alexander Benlian, and Guy Paré. A knowledge development perspective on literature reviews: Validation of a new typology in the is field. Communications of the AIS, 46, 2020.

[70] Sebastian W Schuetz, Paul Benjamin Lowry, Daniel A Pienta, and Jason Bennett Thatcher. The effectiveness of abstract versus concrete fear appeals in information security. Journal of Management Information Systems, 37(3):723–757, 2020.

[71] Michael James Scott, Gheorghita Ghinea, and Nalin Asanka Gamagedara Arachchilage. Assessing the role of conceptual knowledge in an anti-phishing educational game. In 2014 IEEE 14th International Conference on Advanced Learning Technologies, pages 218–218. IEEE, 2014.

[72] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In Proceedings of the 3rd symposium on Usable privacy and security, pages 88–99, 2007.

[73] Mario Silic and Paul Benjamin Lowry. Using design-science based gamification to improve organizational security training and compliance. Journal of Management Information Systems, 37(1):129–161, 2020.

[74] Michael Sorensen. The new face of phishing. https://apwg.org/the-new-face-of-phishing/, 2018. Accessed: 2021-01-13.

[75] Nathalie Stembert, Arne Padmos, Mortaza S Bargh, Sunil Choenni, and Frans Jansen. A study of preventing email (spear) phishing by enabling human intelligence. In 2015 European Intelligence and Security Informatics Conference, pages 113–120. IEEE, 2015.

[76] Simon Stockhardt, Benjamin Reinheimer, Melanie Volkamer, Peter Mayer, Alexandra Kunz, Philipp Rack, and Daniel Lehmann. Teaching phishing-security: which way is best? In IFIP International Conference on ICT Systems Security and Privacy Protection, pages 135–149. Springer, 2016.

[77] Richard H Thaler and Cass R Sunstein. Nudge: Improving decisions about health, wealth, and happiness. Penguin, 2009.

[78] Ke Tian, Steve T. K. Jan, Hang Hu, Danfeng Yao, and Gang Wang. Needle in a Haystack: Tracking Down Elite Phishing Domains in the Wild. In Proceedings of the Internet Measurement Conference 2018, IMC '18, page 429–442, New York, NY, USA, 2018. Association for Computing Machinery.

[79] Paul Van Schaik, Debora Jeske, Joseph Onibokun, Lynne Coventry, Jurjen Jansen, and Petko Kusev. Risk

perceptions of cyber-security and precautionary behaviour. Computers in Human Behavior, 75:547–559, 2017.

[80] Anthony Vance, Jeffrey L Jenkins, Bonnie Brinton Anderson, Daniel K Bjornn, and C Brock Kirwan. Tuning out security warnings: A longitudinal examination of habituation through fMRI, eye tracking, and field experiments. MIS Quarterly, 42(2):355–380, 2018.

[81] Gaurav Varshney, Anjali Sardana, and Ramesh Chandra Joshi. Secret information display based authentication technique towards preventing phishing attacks. In Proceedings of the International Conference on Advances in Computing, Communications and Informatics, pages 602–608, 2012.

[82] Rakesh Verma and Keith Dyer. On the Character of Phishing URLs: Accurate and Robust Statistical Learning Classifiers. In Proceedings of the 5th ACM Conference on Data and Application Security and Privacy, CODASPY '15, page 111–122, New York, NY, USA, 2015. Association for Computing Machinery.

[83] Melanie Volkamer, Karen Renaud, Benjamin Reinheimer, and Alexandra Kunz. User experiences of TORPEDO: tooltip-powered phishing email detection. Computers & Security, 71:100–113, 2017.

[84] Rick Wash. How Experts Detect Phishing Scam Emails. Proceedings of the ACM on Human-Computer Interaction, 4(CSCW2):1–28, 2020.

[85] Rick Wash and Molly M Cooper. Who provides phishing training? facts, stories, and people like me. In Proceedings of the 2018 chi conference on human factors in computing systems, pages 1–12, 2018.

[86] Patrickson Weanquoi, Jaris Johnson, and Jinghua Zhang. Using a Game to Teach About Phishing. In Proceedings of the 18th Annual Conference on Information Technology Education, pages 75–75, 2017.

[87] Markus Weinmann, Christoph Schneider, and Jan Vom Brocke. Digital nudging. Business & Information Systems Engineering, 58(6):433–436, 2016.

[88] Zikai Alex Wen, Zhiqiu Lin, Rowena Chen, and Erik Andersen. What. hack: engaging anti-phishing training through a role-playing phishing simulation game. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pages 1–12, 2019.

[89] Oliver Wiese, Joscha Lausch, Jakob Bode, and Volker Roth. Beware the downgrading of secure electronic mail. In Proceedings of the 8th Workshop on Socio-Technical Aspects in Security and Trust, pages 1–9, 2018.

[90] Ryan Wright, Kent Marett, and Jason Thatcher. Extending Ecommerce Deception Theory to Phishing. In Proceedings of the 35th International Conference on Information Systems, 2014.

[91] Ryan T Wright and Kent Marett. The influence of experiential and dispositional factors in phishing: An empirical investigation of the deceived. Journal of Management Information Systems, 27(1):273–303, 2010.

[92] Min Wu, Robert C Miller, and Simson L Garfinkel. Do security toolbars actually prevent phishing attacks? In Proceedings of the SIGCHI conference on Human Factors in computing systems, pages 601–610, 2006.

[93] Min Wu, Robert C Miller, and Greg Little. Web wallet: preventing phishing attacks by revealing user intentions. In Proceedings of the second symposium on Usable privacy and security, pages 102–113, 2006.

[94] Aiping Xiong, Robert W Proctor, Weining Yang, and Ninghui Li. Embedding training within warnings improves skills of identifying phishing webpages. Human factors, 61(4):577–595, 2019.

[95] Weining Yang, Aiping Xiong, Jing Chen, Robert W Proctor, and Ninghui Li. Use of phishing training to improve security warning compliance: evidence from a field experiment. In Proceedings of the hot topics in science of security: symposium and bootcamp, pages 52–61, 2017.

[96] Huiping Yao and Dongwan Shin. Towards preventing qr code based attacks on android phone using security warnings. In Proceedings of the 8th ACM SIGSAC symposium on Information, computer and communications security, pages 341–346, 2013.

[97] Ka-Ping Yee and Kragen Sitaker. Passpet: convenient password management and phishing protection. In Proceedings of the second symposium on Usable privacy and security, pages 32–43, 2006.

[98] Chuan Yue. Preventing the revealing of online passwords to inappropriate websites with logininspector. In Presented as part of the 26th Large Installation System Administration Conference (LISA), pages 67–81, 2012.

[99] Leah Zhang-Kennedy and Sonia Chiasson. A Systematic Review of Multimedia Tools for Cybersecurity Awareness and Education. ACM Computing Surveys (CSUR), 54(1):1–39, 2021.

[100] Verena Zimmermann and Karen Renaud. The Nudge Puzzle: Matching Nudge Interventions to Cybersecurity Decisions. ACM Transactions on Computer-Human Interaction, 28:7:1 – 7:45, 2021.

Appendix

Database	After search	After exclusion
ACM	270	35
IEEE	869	15
Web of Science	970	25
NDSS/(Euro)USEC	5	2
USENIX Security/SOUPS	8	3
Other	2	2

Table 1: Number of articles included in the literature review before and after applying the exclusion criteria during the screening of title and abstract.

Author	Sample Size	Lab Study	Online Study	Field Study	Survey	Conceptual	Education	Training	Awareness-raising	Design	Email	URL	Website	Authentication	SSL	Other	Pre-Decision	During Decision	Post-Decision	Interactive	Passive	Educational	Non-Educational
		Method					Intervention Category				Attack Vector						Time of Interv.			Activity		Educ.	
Abbasi et al. [2]	509		•						•			•	•					•		•			•
Alnajim & Munro [4]	36	•						•				•	•						•	•		•	
Arachilage et al. [5]	20	•						•									•			•		•	
Baslyman & Chiasson [6]	21	•						•									•			•		•	
Beguin et al. [7]	14	•						•									•			•		•	
Blythe et al. [8]	/					•	•		•								•	•		•		•	
Burns et al. [10]	400			•				•			•	•							•	•		•	
Burns et al. [11]	/					•		•									•			•		•	
Canova et al. [12]	19	•	•					•				•					•			•		•	
Caputo et al. [13]	1,359			•				•			•	•							•	•		•	
Carella et al. [15]	150			•				•									•			•		•	
Cuchta et al. [21]	4,777			•				•			•	•							•	•		•	
De Ryck et al. [22]	/					•				•			•					•			•		•
Dhamija & Tygar [24]	/					•				•				•				•			•		•
Egelman et al. [25]	60	•							•				•					•		•	•		•
Fatima et al. [28]	63	•						•									•			•		•	
Gastellier-Prevost et al. [29]	/					•			•				•					•		•			•
Gokul et al. [17]	8,071		•					•				•					•			•		•	
Greene et al. [30]	ca. 70			•	•			•			•								•		•		•
Hale et al. [32]	/					•		•									•			•		•	
Hale & Gamble [31]	/					•		•									•			•		•	
Herzberg & Jbara [34]	23	•								•		•			•			•			•		•
Herzberg & Margulies [35]	400			•						•				•				•		•			•
Iacono et al. [36]	18		•							•			•					•			•		•
Jakobsson & Myers [38]	/					•				•				•				•			•		•
Jansen & van Schaik [39]	786				•		•		•					•			•				•	•	
Jensen et al. [40]	355			•				•			•	•					•			•		•	
Kirlappos & Sasse [43]	36		•							•			•					•		•		•	
Kumaraguru et al. [47]	4,517			•				•			•	•							•	•		•	
Kumaraguru et al. [44]	515		•					•			•	•							•	•		•	
Kumaraguru et al. [46]	311			•				•			•	•							•	•		•	
Kumaraguru et al. [45]	30	•						•			•	•							•		•	•	
Lastdrager et al. [48]	353	•					•				•	•	•				•			•		•	
Li et al. [49]	20	•								•		•	•					•		?	?	?	?
Lin et al. [50]	22	•								•		•						•			•		•
Marforio et al. [51]	221	•								•								•			•		•
Marsden et al. [52]	11,968		•					•				•							•	•		•	
Miyamoto et al. [54]	23	•								•		•		•				•		•			•
Nicholson et al. [56]	279		•							•	•							•			•		•
Perrault [60]	462				•			•			•	•					•			•		•	
Petelka et al. [61]	701		•						•			•						•		•			•
Reeder et al. [62]	773			•					•					•	•	•		•		•		?	?
Reinheimer et al. [63]	409			•			•					•				•	•			•		•	
Ronda et al. [66]	2,050			•						•				•				•			•	•	
Schechter et al. [68]	67	•							•	•		•	•	•				•		•			•
Schuetz et al. [70]	264		•				•		•								•			•		•	
Scott et al. [71]	/					•		•				•					•			•		•	
Sheng et al. [72]	42	•						•				•					•			•		•	
Silic & Lowry [73]	384			•				•			•	•					•			•		•	
Stembert et al. [75]	24	•						•	•		•	•						•		•	•	•	
Stockhardt et al. [76]	81	•						•				•					•			•		•	
Varshney et al. [81]	/					•				•				•				•			•		•
Volkamer et al. [83]	16			•					•	•		•						•		•		•	
Wash & Cooper [85]	1,945			•			•	•			•	•							•	•		•	
Weanquoi et al. [86]	/					•		•									•			•		•	
Wen et al. [88]	39	•						•			•	•				•	•			•		•	
Wiese et al. [89]	18		•						•	•	•							•		•		•	
Wu et al. [93]	21	•							•	•				•				•		•		•	
Wu et al. [92]	30	•							•	•		•			•			•			•		•
Xiong et al. [94]	639		•					•				•						•		•		•	
Yang et al. [95]	63			•			•		•			•						•		•		•	
Yao & Shin [96]	20	•							•							•		•		•		•	
Yee et al. [97]	/					•				•				•				•		•		•	
Yue et al. [98]	/					•			•					•				•		•			•
Sum (N=64)		20	12	16	3	13	7	31	17	20	17	33	10	12	4	4	23	31	11	48	16	43	19

Table 2: Results of the literature review, sorted alphabetically by first author.

Investigating Web Service Account Remediation Advice

Lorenzo Neil
North Carolina State University

Elijah Bouma-Sims
North Carolina State University

Evan Lafontaine
North Carolina State University

Yasemin Acar
Max Planck Institute for Security and Privacy

Bradley Reaves
North Carolina State University

Abstract

Online web services are susceptible to account compromises where adversaries gain access to a user's account. Once compromised, an account must be restored to its pre-compromise state in a process we term "account remediation." Account remediation is a technically complex process that in most cases is left to the user, though some web services provide guidance to users through help documentation. The quality of this account remediation advice is of paramount importance in assisting victims of account compromise, yet it is unclear if this advice is complete or suitable. In this paper, we analyze account remediation advice from 57 popular U.S.-based web services. We identify five key phases of account remediation, use this five-phase model to develop a codebook of account remediation advice, then analyze topic coverage. We find that only 39% of the web services studied provided advice for all phases of account remediation. We also find that highly-ranked websites and sites with a previously disclosed data breach have more complete coverage than other sites. Our findings show that account remediation should be more carefully and systematically considered by service providers, security researchers, and consumer advocates, and our detailed analysis will aid in creating better guidelines for users and services.

1 Introduction

Online web services allow people to create accounts that store information and communicate with others. Compromises of these accounts are a pervasive problem, with billions of accounts being compromised in 2019 alone [21]. Account compromises allow the attacker to steal service, surveil the activities of the victim, abuse the system, or otherwise compromise the confidentiality, integrity, or availability of the account. When compromised, an account must be re-secured in a process we term *account remediation*. In this work, we determine that there are five key phases for account remediation. In order, these are: detecting the compromise, recovering access to the account, limiting access by the attacker, restoring the account state and associated data to the pre-compromise state, and taking action to prevent future compromises.

After having accounts compromised, the authors discovered first-hand how technically complex and frustrating the task of account remediation can be. We found anecdotally that help documentation provided by web services differs drastically in terms of completeness. When documentation on remediating compromises is lacking, it is much more difficult for users, even technically-savvy users, to remediate a compromise. Therefore, the advice given by web services to help users remediate their accounts is of critical importance. We realized that not only is the advice given to users critical for navigating the process correctly and effectively, but the advice also acts as a proxy for understanding how the organization responsible for creating it views the process.

In this paper, we make the following contributions:

- **Model Account Remediation:** We develop a five-phase model to capture each phase of account remediation, from initial compromise discovery to remediation. We then use this five-phase model to fully represent the range of activities a user may engage in during account remediation in a qualitative codebook.
- **Characterize Webservice Account Remediation Advice:** We use our codebook to evaluate the account remediation advice of 57 popular web services in the United States, providing a window into the resources available to users as well as acting as an implicit measure of web services' own understanding of the issue. We find this advice is sparse and underspecified, especially when we

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

examine activities unique to account remediation. For example, fewer than half of the services studied provide *any* guidance to limit further access by an attacker.

- **Broad Trends and Recommendations:** We find that average phase coverage is higher for services that either are very popular or that have a previously disclosed data breach. We also provide recommendations for web service owners and future researchers.

We note that *account recovery,* defined as the process of restoring a legitimate user's access to an account if credentials are lost or changes, has received substantial research coverage, as we discuss in Section 2. However, account recovery is only a single phase of account remediation. Areas such as limiting an account's access and restoring an account's original state are crucial for account remediation, but have received little research attention.

2 Related Work

A user's mental model on security ultimately informs their security decisions with their devices and online services [12]. Prior research has focused on the user's security mental model [5,29] and how they interpret security advice and warnings [1]. Improving a user's basic knowledge in security limits the chances of their online services being compromised [5], though users may reject the advice if it presents a poor cost-to-benefit ratio or it threatens their privacy [15, 29, 29, 30]. Previous work has found that it is hard for end users, and even experts to prioritize security advice [30]. User advice can cover all five phases of account remediation, though a significant body of work has focused on detecting compromise and account recovery.

Many account compromises stem from stolen credentials. Prior work has measured how the risk of stolen credentials varies between phishing, malware, or data breaches and predicts the chances for total online account takeover from stolen credentials [24, 27, 34]. Billions of stolen usernames and passwords are also widely available in underground forums [25, 35–37]; these data sets have been used to create systems that alert users if their usernames or passwords are vulnerable and have been publicly exposed [25, 35, 37]. Other work on detecting compromised accounts [33] focused on building models to represent normal account behavior and then using that behavior to analyze current account behavior for anomalies or unusual activity [6, 8, 19, 31]. Recent work has investigated whether users are informed about data breaches, how they feel about them, and whether they have taken or plan on taking action [22]. In our work, we go beyond compromise discovery and account recovery, also focusing on remediating harm to the compromised accounts.

Account recovery mechanisms restore access to an account after credentials are lost or changed by an attacker after a compromise. Virtually all widely used password recovery mechanisms, including secret questions and e-mail reset links, have well-understood vulnerabilities and deployment limitations [26]. Many major webmail providers employ security questions that can be solved through data mining, are easily guessable, or have low memorability over time [3, 32]. Prior work on account recovery mechanisms investigated different authentication schemes [4] and password reset strategies [16]. Password recovery schemes may also be vulnerable to man-in-the-middle (MitM) attacks [13, 14]. Compromise detection and account recovery have both been widely studied topics, yet to the best of our knowledge, we are the first to study account *remediation* from a holistic perspective.

3 Methods

In this section, we describe our methods (see Figure 1): codebook development (3.1), account remediation model creation (3.2), ensuring inter-rater reliability among coders (3.3), coding account remediation advice from 57 web services (3.4), and our analysis of differences in the coverage of account remediation advice among web services based on their popularity and disclosure of data breaches (3.5).

3.1 Codebook Development

Three authors created the codebook deductively based on nine popular web services' account remediation advice, inductively informed by authors' personal and professional experience with account remediation, and existing research on account recovery, data breach notification and behavior, and authentication. We first annotated nine popular web services' account remediation advice, [1] then iteratively built and revised our codebook and operationalized the codes. We finalized our codebook when we were able to unambiguously apply it to assess account remediation advice for the initial nine web services. This was evidenced by high agreement when applying the codebook (Krippendorff's Alpha > 0.75 for all three coders for independent coding [9, 10]). In line with recommendations for qualitative coding, we used this score not only to assess our level of agreement, but also to investigate where and how we disagreed [2, 23]. If that coefficient was not met when we compared our codes, we used it as an opportunity to better define and disambiguate codes, as well as discuss what causes confusion or disagreement.

The final codebook contains five top-level codes, *compromise discovery*, *account recovery*, *limiting access*, *service restoration*, and *prevention*, which we call the five phases of account remediation, as well as sub-codes that represent concrete advice, For example, in *prevention*, we have a sub-code "enable 2FA", which describes advice to enable 2-FA for an account to prevent a *future* compromise.

[1] Facebook, Netflix, Skype, Spotify, Twitter, LinkedIn, Google, Yelp, Walmart

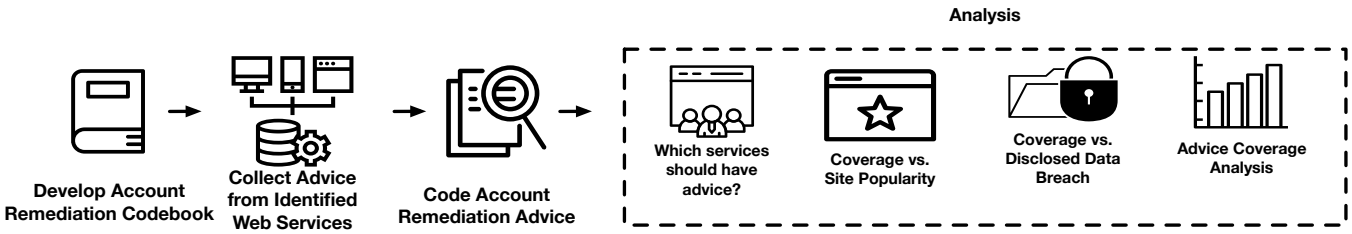


Figure 1: Methodology: codebook and model development, data collection and analysis.

3.2 Account Remediation Model

We explain the account remediation process as five phases of account remediation: *compromise discovery*, *account recovery*, *limiting access*, *service restoration*, and *prevention*, corresponding to our codebook's top-level categories.

Compromise discovery describes a user observing suspicious activity from their account or service that indicates a possible compromise, for example: "If you notice unfamiliar activity on your Google Account, someone else might be using it without your permission." (Google).

Account recovery describes the process for users to regain access to their account after losing access to it or having it compromised. We differentiate account remediation from account recovery in the sense that account recovery is only one phase in the account remediation process. An example of advice for *account recovery* is: "Change your password or send yourself a password reset email." (Instagram)

Limiting access describes preventing current and future unauthorized access from adversaries, for example: "Sign out of all devices connected to your account unless you believe your device has been stolen." (Netflix)

Service restoration describes restoring an account's original settings, content, or state before a compromise. An example of advice for *service restoration* is: "After signing in, you'll want to review the recent activity on your account." (Microsoft)

Prevention describes preventing future compromises by taking steps to further secure an account, like "Never click suspicious links, even if they appear to come from a friend or a company you know" (Facebook).

While advice coverage was not uniform across services, we found that the concept of the top-level categories (our five phases) was present across services. While we theorize that these five phases conceptualize account remediation in general, we do not imply that each specific subcode in each phase has to be covered by all services to provide complete advice, as service offerings may differ. We established this model to account for a wide range of advice and describe the majority of account remediation steps.

3.3 Training and Reliability

Following the development of the codebook, the main author trained two supporting coders on the nine initial web services, again measuring inter-rater reliability to pinpoint and resolve disagreement and to determine the successful conclusion of the training phase. The agreed-upon coding by the three codebook developers was used as ground-truth for training, and once Krippendorff's Alpha consistently exceeded 0.75, we considered the new coders competent to apply the codebook [10].

After the training phase concluded, the supporting coders then coded the rest of the web services individually. The primary coder independently double-coded a select subset of web services from each supporting coder, usually those that had been subjectively the hardest to code. After each week of independent coding, the primary coder met with each supporting coder separately to resolve disagreements, errors, and confusion, as well as to make sure that coding strategies did not diverge over time.

3.4 Collecting Advice from Web Services

In this section, we explain our process for web service selection, how we collect and store the advice, and how we established groups of web services for research questions.

Service Selection Criteria: We referred to two lists generated from the Tranco Website Ranking Service [28] to identify web services of interest. The lists were generated on March 31, 2020 and August 18, 2020. Using these Tranco lists as a reference, we examined web services that were U.S.-based, allowed user online account creation, and provided publicly available account remediation advice. We chose U.S.-based web services since all authors are fluent in English. We excluded adult-content web services from the study, as our research was performed on computers owned by a public university. Finally, we excluded services that were unreachable at the time of data collection.

Finding Advice: To ensure the totality of advice collection, we collected account remediation advice from web services by both manually browsing their help pages and through search queries on the website and Google. When navigating the web

service, we searched both the help center sections and security settings (if available). We queried the help center with the template phrases: "My account was compromised" and "My account was hacked". Once we found a web service's main page for account remediation advice, we also collected every relevant link mentioned on that page for account remediation. We further extended our collection of advice by Google search querying for any account remediation advice from the target web service based on text snippets we found on advice sites, our own experiences, and anticipating the spectrum of possible user queries. Our Google search queries were the following: "My [web service] account was compromised" and "My [web service] account was hacked". We added any new account remediation advice that was not found when navigating the web service. This multi-step process ensured that we identified all relevant account remediation advice from a web service. We note that many large companies have separate web services served by the same account management; one example is Google and YouTube. In such cases, we only include an advice policy once.

Content Exclusion Criteria: For our analysis, not all information is appropriately considered account remediation advice. For example, we do not consider advice for accounts that were suspended due to actions of the user or suspensions that were self-inflicted. Secondly, we did not include advice within forums or posts by other users on the service or on third-party sites, because such information may be inaccurate, outdated at the time of collection, or from an untrustworthy source. We also exclude advice documents when they consisted *solely and entirely* of a suggestion to contact the service.

We also only collect advice available without requiring a logged-in web service account to replicate the process a user would take if they could not access their compromised account and needed guidance. This strategy also allowed us to collect all relevant advice regardless of the login status. After our initial data collection, we observed that financial services and universities had been almost entirely excluded by this strict criteria. Owing to the importance of these two industries as targets of compromise, we revisited these services to collect publicly available remediation advice. Out of an abundance of caution, in Section 4.3 we include results with and without the financial service and university data.

Collected Datasets: We divide our collected data into two groups, shown in detailed tables in the Appendix. Both groups account for 57 total web services. The *very popular* web services dataset consisted of the top 31 web services (as ranked by Tranco) that were U.S.-based and offered account remediation advice. To this dataset, we added one additional service (Yelp) slightly outside of the Top 31 that had been chosen arbitrarily as a case study during codebook creation. We note that after filtering by our criteria and excluding combined web properties from the list (e.g., Google and YouTube) our first 31 services span from Google (ranked #1) to Walmart (ranked #184), with our last service (Yelp) ranked 209 at the time of data collection. Therefore, this group consists of 32 web services and we will refer to them as our *very popular* set of web services throughout the paper. We explain in Section 3.5 how we define popularity.

Our second dataset, termed the *less popular* web services dataset, consisted of a random selection of 25 services meeting our full criteria with a Tranco rank in the range of 500–1000. Initially, we aimed to collect advice from 32 web services in this range in order to have two equal sets of web services. However, upon coding these web services in the full study, the coders had trouble coding the advice specifically in regards to advice from the phase compromise discovery. The confusion came from the fact that it was hard to differentiate whether advice to discover a compromised account was either solely billing/financial issues or actually other codes related to compromise discovery. Due to this confusion, we decided to discard banking web services in this group of web services, which left us with 25 *less popular* services, as we will refer to them throughout the paper. This specific range was chosen to select a group of web services that were not obscure but was also noticeably different from the very popular web services ranked at the top. Rankings like Tranco in general are rarely linear in correlation with the phenomena measured (or implied). For example, consider the case of Youtube, Netflix, and Crunchyroll. Youtube and Netflix were ranked 3^{rd} and 9^{th} respectively, while Crunchyroll was ranked 837^{th}. Though Youtube is ranked 3 times higher than Netflix, it is unlikely that YouTube has three times the resources for security than Netflix; nor is it likely the case that Youtube has nearly a three-orders of magnitude larger security budget than Crunchyroll. Consequently, to see if site popularity has an effect on remediation advice coverage, we choose to look at group distances between the rough equivalence classes formed by the broad rank range.

Recording Existence of Account Remediation Advice: Using the same selection and exclusion criteria, we analyze all web services that were ranked between 500- -1,000 on Tranco [28] for existence of publicly available account remediation advice. We are not coding web services here; we simply check if web services provide public account remediation advice. Therefore, we examine all web services in this range, not just web services with account remediation advice. Once we calculated how many web services fit our selection criteria and provided public account remediation advice, we divided that number by the total number of web services that fit our selection criteria.

We perform this method on two different data sets, each data set however consists of web services ranked between 500- -1,000. Both data sets consisted of web services that were U.S.-based and allowed for account creation. The difference is that the first data set will also include financial or university-based web services that were ranked between 500- -1,000. We refer to this data set throughout the paper as the *include financial/university* services group. The second

data set is identical but excludes financial or university-based web services ranked between 500- -1,000. We define this data set throughout the paper as the *exclude financial/university* services group. We include two data sets since we cannot confirm if financial-based or university web services provide different account remediation advice to users with a login or belonging to that community. Since our criteria were to only collect advice that was publicly available without a login, we separate our findings for this question. These results will be shown in Section 4.3.

Storing Advice: When we found all relevant account remediation advice from a web service, we saved PDF versions of the web pages and stored them for analysis. This helped ensure we had a static dataset that did not change as we were coding. This also allowed us to code web services both collectively or individually by analyzing similar PDFs for web service's account remediation advice.

Coding: After the training phase was complete, each new coder coded 22 web services (totaling 44 more web services). Each coder coded the PDF pages from the web service's advice with Nvivo, in increments of five to nine web services at a time. Once each week, the first author met with both coders separately to go over the overlapping coding results and resolve confusion or disagreements about the coding results. Each coder then corrected their codes or added codes that they missed. Our coding results are in Sections 4.1 and 4.2.

3.5 Differences in Coverage of Account Remediation Advice

In this paper, we seek to understand whether there are significant differences in the coverage of account remediation advice between *very popular* web services and *less popular* web services, and whether there are significant differences in the coverage of account remediation advice between web services with a disclosed data breach and web services without a disclosed data breach. To address these questions, we need to operationalize aspects of these questions, including coverage, popularity, breach history, and group differences. This subsection presents the methods we use for each of these issues.

We operationalize the coverage of account remediation advice as a web service covering all five phases of account remediation in their advice. The range of the coverage of advice is measured from one phase coverage up to five phases coverage. For example, if a web service gives advice that covers only compromise discovery, account recovery, and limiting access, the coverage of the advice for that web service will be a three since it mentioned advice from three phases. We define the coverage of advice for account remediation in this manner because every phase for account remediation is important in successfully remediating a compromised account. However, not every individual code in every phase will be relevant or important for every web service. For example, codes for advice on noticing billing/finance issues will not be relevant for web services that do not handle money transactions or store financial information. Also, web services that do not give users the functionality to install third-party applications will not need advice on how to remove potentially malicious third-party applications. For this reason, if a web service has advice that mentions at least one code from a given phase, that phase will be counted to the coverage of account remediation advice for that web service. While this may overestimate a service's advice (i.e., coverage does not imply a high quality of advice), we can confidently assess services with low coverage and services with high coverage of advice.

We define the popularity of a web service by its ranking on the Tranco Website Ranking Service [28]. This ranking service was developed mainly for research purposes and consists of data from many ranking services over a period of 30 days. Tranco lists web services based on their popularity. The top 32 ranked web services we analyze are at the very top of this list, called here the *very popular* group of services. Lower ranked web services on the list such as the 25 randomly sampled web services in the 500-1,000 range are the *less popular* group of services. Our results for comparing the differences in coverage between *very popular* web services and *less popular* web services are shown in Section 4.4.

We operationalize "public disclosure of data breaches" by using a well-known database maintained by Troy Hunt on his website "haveibeenpwned" [17]. Haveibeenpwned consists of a database of publicly disclosed data breach incidents that have been consolidated and displayed on the website. The database also contains hundreds of database dumps and paste bins containing billions of leaked account credentials. Users then can query this website to search if their credentials such as their emails, usernames, or passwords have been compromised or "pwned." Users can also check an overview of web services that haveibeenpwned has listed as being breached, and sign up for breach notification. When we define web services to have publicly disclosed a data breach, we refer to web services that are listed on haveibeenpwned; the *data breach disclosed* group contains 16 web services. The remaining 41 web services that were not mentioned in the breached list of web services [17] make up our *non-data breach disclosed* group. Our results for comparing the differences in coverage between *data breach disclosed* and *non-data breach disclosed* web services are shown in Section 4.5.

In order to statistically evaluate the differences in our two research questions, we perform a Mann-Whitney U Test for both questions. Specifically, we investigate if the means of the distribution of the number of phases within the groups involved in the research questions is significant in difference. Using the Mann-Whitney U scores, we then calculate the magnitude in differences between each group of web service's coverage of advice by calculating their respective ef-

fect size [11,20]. This effect size is also quantified in Cohen's confidence interval r [7]. We follow the interpretations as guidelines provided by Fritz [11], which describe r = 0.1 as "small", r = 0.3 as "medium", and r = 0.5 as "large". The Mann Whitney U Test and other related statistical measures were performed with SPSS software [18]. We then used these results to calculate the effect size [11].

3.6 Limitations

As with any study that involves qualitative coding, this study is subject to the authors' biases, as well as possible differences in coding strategies between coders. We tried to reasonably address these in our investigation by having coders with diverse research backgrounds on our team to allow multiple perspectives to inform the creation of our codebook, and, eventually, the five phase model of account remediation. We also diligently refined our codebook and the codes' explanations in order to allow independent coders to arrive at similar assessments, and regularly controlled for divergent strategies, discussed differences and resolved disagreements.

Additionally, due to the nature of our study, we cannot provide ground truth about the differences in the coverage of account remediation advice between different groups of web services. Our definition in the coverage of advice does not take into account the length or depth of the advice, rather a metric for how many phases in account remediation it covers. We also do not provide ground truth for the applicability of all of the codes in our codebook to web services. Most of the codes in our codebook represent advice that can be broadly applied to all web services. However, some codes that we developed during our codebook development like "observe billing" or "finance issues" or "observe a third party account connected" do not apply to all web services. Therefore, we explain the results for specific codes like this with the caveat that they may not be broadly applicable to all web services.

We only collect advice from web services when it was publicly available without an account login. Some web services may provide additional account remediation advice once a user is logged in. We collected advice in this manner to replicate the process of finding account remediation advice, in the case where the account owner cannot access their account. For our coding results in Sections 4.1, 4.2, 4.4, and 4.5, we only include web services that provide publicly available account remediation. In Section 4.3, we include two versions of results in which we exclude web services that may provide additional account remediation advice given an account login.

Lastly, we only included U.S-based web services in this study. We wanted to ensure that all coders could fully interpret and code the web services we selected for this work. Since the only language that every author could fluently speak is English, we limited ourselves to U.S.-based web services.

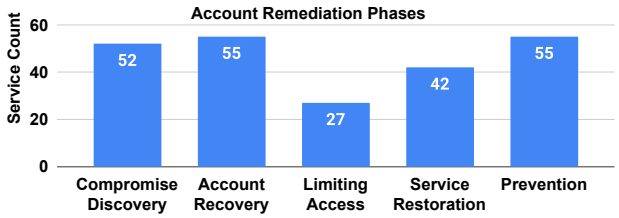


Figure 2: Bar graph of all account remediation phases among web services. Limiting Access advice is mentioned in less than half of the web services we analyzed. Service Restoration advice was mentioned in 74% of the web services. All other phases were mentioned by at least 90% of the web services we analyzed.

4 Results

In this section, we discuss the results of our codes and implications behind the results. In Section 4.1, we provide the overall coverage of the phases for account remediation advice from the web services. In Section 4.2, we look at each phase individually and examine the coverage of their respective codes within the web services. In Section 4.3, we report how many of bottom 500 ranked web services provided users with publicly available account remediation advice. We present this report with the inclusion of financial web services and university web services and also without financial web services and university web services. In Section 4.4, we present our results for investigating the differences in the coverage of account remediation advice between *very popular* and *less popular* web services. Similarly in Section 4.5, we present our results for investigating the differences in the coverage of account remediation advice between *data breach disclosed* services and *non-data breach disclosed* services.

4.1 Overall Phase Coverage

Sections 4.1 and 4.2 reflect results from coding all 57 web services. Advice for compromise discovery, account recovery, and prevention was mentioned by 91%, 96%, and 96% of all web services, respectively. These were the only phases that were covered in at least 80% of account remediation advice from web services. On the other hand, advice for limiting access was mentioned by 46% of web services and advice for service restoration was mentioned by 75% of web services. Figure 2 represents web service counts for every phase in the account remediation model. The service count in the graph indicates how many web services mentioned at least one code from a specific phase.

The phases of compromise discovery, account recovery, and prevention are not only widely addressed by most web services, but also represent areas that have been heavily re-

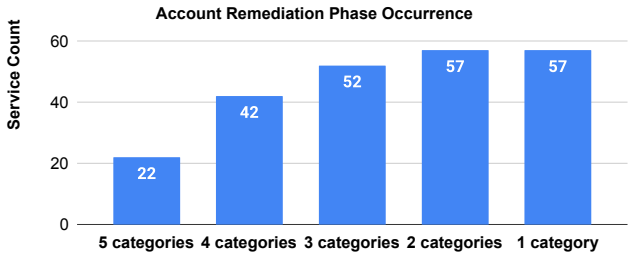


Figure 3: How many web services mentioned at least n amount of web services, where n is either at least 5,4,3,2,or 1 phase. Only 39% web services gave advice for all five phases.

searched by the security community. These three phases, however, do not fully cover the process of account remediation. Limiting the access of an account and restoring an account's original settings are fundamental for account remediation. Without it, the account remediation process is not complete, and a compromised account may still remain vulnerable. Still, more than half of the web services we investigated did not mention any advice for limiting an attacker's access.

Out of the total 57 web services we analyzed for account remediation advice, only 39% managed to mention advice from all five phases. 74% of web services mentioned at least four account remediation phases. 91% of web services mentioned at least three account remediation phases. Lastly, all 57 web services mentioned at least two account remediation phases. Figure 3 shows these results from coding all 57 web services.

The consequences of these results require careful consideration. On the one hand, our results for security advice most unique to account remediation (limiting access and service restoration) would seem to indicate that web services are neglecting these two phases. On the other hand, while we believe our model is sufficiently general to capture the account remediation process, there may be cases where it is not necessary to cover all five phases explicitly. Consider a hypothetical service that recommends completing the account recovery process, and it happens to log out all logged-in sessions. The service's advice may not reflect any limiting access content because it is automatically handled. Without ground-truth knowledge about each web service's internals, it is difficult to determine which case applies to a particular web service. Taken together, it is clear that future work should determine if remediation phase coverage is low because it is neglected or if it is simply not necessary.

4.2 Content Analysis by Phase

Compromise discovery: Compromise Discovery involves observing activity from an account or service that indicates a possible compromise. Our results for the compromise discovery codes are shown in Figure 4. *Only 11% of the codes in this phase were covered by at least half of the web services.*

Advice for discovering unauthorized or suspicious activity was recorded in 68% of the web services. This was the only advice in compromise discovery however that was mentioned in at least half of the web services. A possible reason for this could be that all of the advice in this phase can be related to unauthorized or suspicious activity, and the code itself is much less specific compared to other codes in this phase. This is a broad interpretation of compromise discovery since there are multiple methods of compromise discovery.

Advice to discover an email change or password change was mentioned in 12%, and 21% of web services, respectively. The majority, if not all, of web services with account creation store a user's email address and password and allow users to change them as well. Observing that either of these identifiers changed within an account is a strong indication of a possible compromise. Still, even the union of the coverage of advice for discovering a changed password and changed email address reached no more than 33% of web services we investigated. This is a clear oversight of advice coverage on the part of web services.

Advice noticing an explicit notification and observing unauthorized logins was mentioned in 30% and 35% of the web serviced we investigated, respectively. We wanted to code advice for users discovering account compromises from explicit notifications from the service, or by observing unauthorized logins on their accounts. From this, we also concluded that users could observe unauthorized logins due to an explicit service notification, or by examining their account as well. Therefore, we created a code for noticing explicit service notifications about a compromise and a code for observing unauthorized logins that includes coverage from the explicit service notification code, while not being exclusive to it. With these results, we present the caveat that we do not confirm if all web services give users the functionality to observe log-ins on their accounts. Therefore, the results for our code "observe an unauthorized login" may not be broadly applied to all web services.

Advice to discover a social media/third party account connected and billing/finance issues were mentioned in only 5% and 35% of web services, respectively. While these results do reflect low coverage across web services, we can not confirm how many web services in our study implement billing or finances into their functionality for users. We also can not confirm if all web services in our study allow users to connect a social media or third-party account to their main account.

We look to our results in coding limiting access advice later in this section and compare the results of the code "Remove third party access." This specific code, "Remove third party access", was mentioned in 18% of web services. The difference in coverage between this code and our code in this category, "social media/third party account connected," shows that at least 12% of web services that allow users to connect a

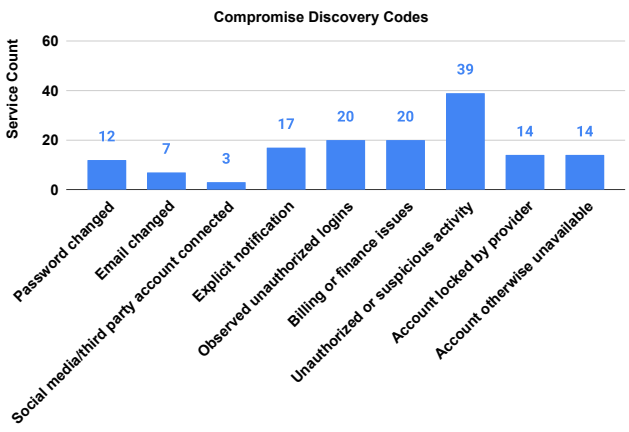


Figure 4: Bar graph of Compromise Discovery codes among web services. Unauthorized or suspicious activity was the highest covered code with 39 web services. No other code was mentioned in more than half of the web services.

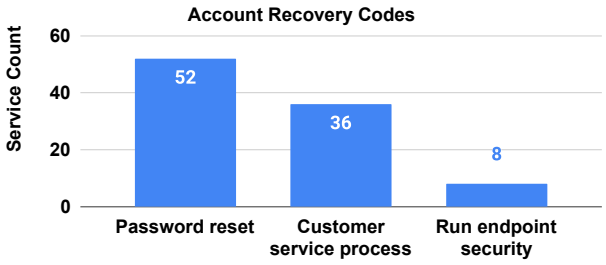


Figure 5: Bar graph of Account Recovery codes among web services. Password reset was mentioned by 91% of services and customer service support was mentioned by 63% of services.

social media or third account are not advising users to notice a new social media or third-party account when discovering a compromise.

Overall, compromise discovery advice was sparsely covered. Only one code in this phase was covered by at least half of the web services. Most of the codes in this phase can either be broadly applied or covered at a higher usage given other results we recorded in other phases. Most of the advice in this phase is also cheap in implementation but important to discovering a compromised account. Web services have much room for improvement in their coverage of compromise discovery advice.

Account recovery: Account recovery provides a means for users to recover their account after losing access to it or having it compromised. Our results for coding this phase are shown in Figure 5. *66% of the codes from this phase were covered in at least half of the web services.* This phase is highly covered by web services and continues to be prioritized, even as a means to remediate compromised accounts.

Advice to initiate a password reset or to change a password was covered in 91% of web services. This advice was also the highest covered code out of all phases in this study. It was the most common method for advising users to recover their compromised accounts.

Advice to advise users to engage in customer service to recover a compromised account was covered by 63% of web services. Some services require contacting customer service for account recovery processes. Customer service for account recovery involves assisting users in recovering a compromised account with a guided process or interaction with a service client. This is different from other customer service processes that services may offer outside of account recovery. While we recognize this advice was not covered universally among web services, it may not be reasonable to have users go through customer service every time to recover their account or reset their password. However, keeping customer service as an optional route may be more beneficial to users.

Advice to reset passwords and to engage in customer service to recover an account were both covered in over half of the web services. These results can imply that not only is account recovery prioritized in account remediation advice, but mainly in the forms of password reset advice and customer service support

We observed advice for running endpoint security to recover an account was only covered in 14% of web services. The low service count could be the result of authors of account remediation advice not considering endpoint security. Also, correctly running anti-virus software is highly technical and possibly beyond the reach of most users. It might be unclear to the extent of how much antivirus or other harm remediation measures help remediate online account compromise. This can imply that web services may not view endpoint security options as a viable solution or prioritize it for account recovery purposes.

Limiting access: Limiting account access is defined as preventing current and future unauthorized access by adversaries. *Limiting Access advice was the lowest covered phase in the study, reaching only 47% of web services.* Less than half of the web services in our study advised users to manage the access of their account, and thus not prioritizing an important step in account remediation. Advice for limiting an account's access includes signing out of instances of an account, reviewing active sessions, and removing access from third-party applications. The results for coding this phase are shown in Figure 6.

Advice for signing out of an individual instance or all instances of an account were covered by only 26% and 14% of web services, respectively. All services allow users to sign out of an account and many allow to sign out of multiple account instances, yet the union of these two codes was only covered

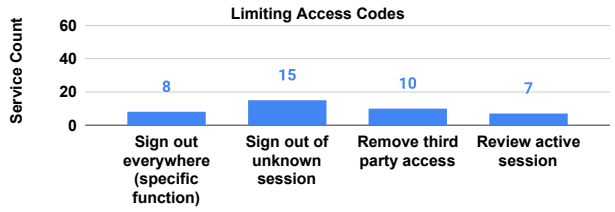


Figure 6: Bar graph of Limiting Access codes among web services. No single code was mentioned in more than a third of the web services.

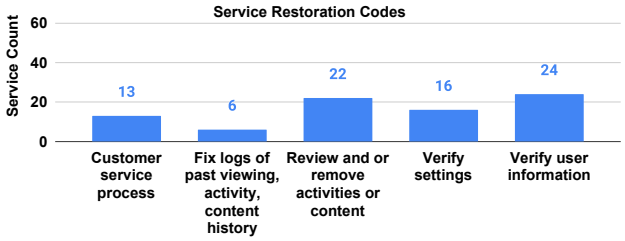


Figure 7: Bar graph of Service Restoration codes among web services. No single code was mentioned in more than 42% of web services.

by 40% of web services. This coverage is insufficient given that all web services allow users to sign out of an instance or multiple instances of their account and it is an important step in managing the access of an account.

Advice for reviewing active sessions also was represented with a code that was only present in 12% of web services. We also record this finding with the caveat that we lack ground truth for how many web services provide users the ability to check for active sessions of their account. However, we explain in Section 5, why we recommend this functionality be implemented in web services and then provided in account remediation advice.

Advice for removing third party access was only present in 18% of web services we investigated. This is important to note since advice for discovering a new social media or third-party account connected to an account in the compromise discovery category was only mentioned in 5% of web services. All of the advice in this phase is underwhelmingly covered given its importance to secure the access of a compromised account.

Service restoration: Service Restoration advice involves restoring an account's original settings or information to how it was before the compromise. *74% of web services mentioned advice for service restoration, yet none of the specific codes in service restoration were covered by at least half of the services.* The results for coding this phase are shown in Figure 7. Advice from this phase is also insufficient in coverage among web services.

Advice for verifying user information, verifying account settings, and reviewing and/or removing activities or content were each recorded in 42%, 28%, and 39% of web services, respectively. These are extremely low percentages for advice that should apply to most, if not all, of the web services we analyzed. All web services in this study store information about the user, settings for the user, and activity by the user. Therefore, there should be advice to verify all of this information. Yet, none of the codes that represent this advice are mentioned beyond 42% of web services investigated.

Lastly, advice to seek customer service support in this phase received a low percentage: 23% of web services. This percentage differs significantly in coverage than the service count for customer service support for account recovery which was mentioned in 63% of web services. This could imply that most services are more likely to prioritize customer service support advice for account recovery, or they do not prioritize customer service for service restoration purposes.

Prevention: Prevention is defined as taking further steps to further secure an account. *Out of the total 11 codes in Prevention, four were represented in at least 60% of the web services investigated.* This category also held six of the top ten most covered codes in the codebook (strong password advice, secure email advice, enable 2FA, check/modify related accounts, enable endpoint security options, and keep software updated). Results for coding this phase are presented in Figure 8.

Advice to maintain strong passwords was the highest mentioned code in this category with 88% coverage. This was the second individual highest covered code right behind the advice to initiate a password reset to recover an account (91% coverage). This means that advice for password security amounted to the two highest codes and therefore the highest coverage out of any advice for account remediation. This could be a result of the vast industry and academic work on password security. It could also mean that web services believe strong password advice is very crucial to account remediation.

Advice on securing emails, enabling two-Factor Authentication, and checking or modifying related accounts was covered in 72%, 70%, and 61% of web services, respectively. Similar to strong password advice, secure email advice and two-factor authentication advice also represent areas that are heavily researched by the research community and are popular among web services.

Running endpoint security options and keeping software up to date advice were both mentioned in 47% of web services investigated. Interestingly, the coverage in this phase for running endpoint security was significantly higher than advice for running endpoint security for account recovery (14%). This shows authors of advice for account remediation were more likely to advise users to run endpoint security options to prevent an account compromise instead of recovering an account from compromise. However, given that it is unclear

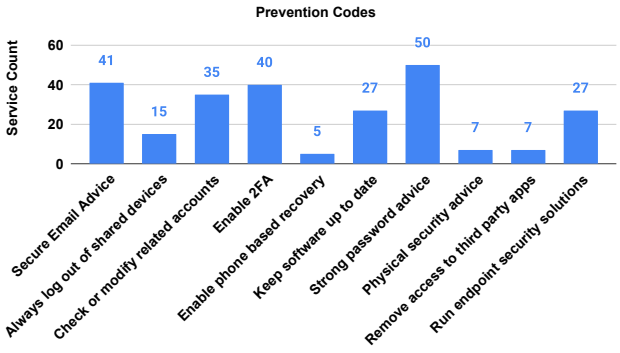


Figure 8: Bar graph of Prevention codes among web services. Four out of 11 codes were mentioned in at least 60% of web services and strong password advice was mentioned in 88% of web services.

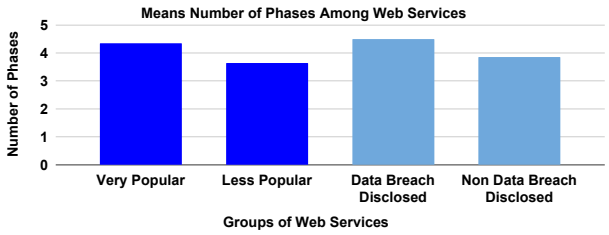


Figure 9: Graph of the mean number of phases covered in account remediation by all experimental groups. *Very popular* web services had a higher mean count of phases mentioned in their account remediation advice than *less popular* web services. *Data breach disclosed* services had a higher mean count of phases mentioned in their account remediation advice than *non-data breach disclosed* web services.

how effective running endpoint security options are towards recovering a compromised account, it is also unclear as to how effective it is in preventing a future compromise.

Notably, prevention advice generally focused on shifting responsibility to other services or the user. While not explicitly coded for, very few services discussed reporting breaches or security flaws in their own service. For example, Netflix states that "If [users] believe [they've] found a security vulnerability on a Netflix property or app, we strongly encourage [them] to inform [Netflix] as quickly as possible and to not disclose the vulnerability publicly until it is fixed." In the worst case, Fandom.com prefaces its prevention advice with the statement that "there is a possibility that if your account is hacked you will need to create a new account" and implies that security is solely the responsibility of the user.

4.3 85% of Web Services did not provide Account Remediation Advice

In our *include financial/university* web service data set, 220 web services allowed users to create public accounts and were U.S.-based. *Of these 220 web services, only 15% of these web services gave publicly available account remediation advice.* In our *exclude financial/university* web service data set, 195 web services allowed users to create a public accounts and were U.S.-based. *Of these 195 web services, only 12% of these web services gave publicly available account remediation advice.*

The majority of web services in our study that were U.S.-based and allowed for user account creation did not provide users with public advice for account remediation. This is alarming since we made sure to only collect account remediation advice from a web service if the advice was publicly available and did not require users to log in. A user with a compromised online account needs to have access to such advice even if they cannot access their account. If this advice is not made publicly available, let alone created at all, then users are left with significantly less help in successfully remediating their compromised accounts.

4.4 Coverage of Advice versus Popularity

In this section, we give our results from investigating the differences in the coverage of account remediation advice between *very popular* web services and *less popular* web services. As stated in Section 3.5, we define the coverage of account remediation advice as the number of account remediation phases that are discussed by a web service.

Our objective is to see if there are differences in the number of phases covered within account remediation advice for web services of vastly different popularity. We performed a Mann-Whitney U Test in which we define the following null hypothesis: the distribution of the number of phases mentioned in account remediation advice is similar across *very popular* web services and *less popular* web services. We perform this test to discover if the number of phases between the two groups of web services is significant in difference.

The mean number of phases mentioned by *very popular* web services was 4.3 with a standard deviation of 0.90. While the mean number of phases mentioned by *less popular* web services was 3.6 with a standard deviation of 0.90. Using a Mann-Whitney U test, we find a statistically significant difference in the mean number of phases covered by the two groups (U = 224, z = -2.994, $p = 0.003$). Using these test scores, we calculate an effect size $r = 0.397$, which is considered to be a "medium" effect size [11, 20].

It is plausible that *very popular* web services have more incentive to provide users with account remediation advice since they have more users creating accounts than less popular web services. Not only would they have more users, but there

may also be a higher importance or usage of accounts with very popular web services. However, there are important web services that are not *very popular*, but are likely to also provide extensive account remediation advice. Financial and banking web services are also important to users, and compromised accounts from these web services can impact a user's finances or potentially compromise their identity. Many banks provide both advice for account remediation and identity theft and also give users resources to contact for further assistance.

4.5 Coverage of Advice versus Disclosed Data Breach

In this section, we show the differences in the coverage of account remediation advice between *data breach disclosed* web services and *non-data breach disclosed* web services.

Our objective is to see if there are differences in the number of phases covered within account remediation advice for web services that have or have not publicly disclosed a data breach. We performed a Mann-Whitney U Test in which we define the following null hypothesis: the distribution of the number of phases mentioned in account remediation advice is similar across *data breach disclosed* web services and *non-data breach disclosed* web services. We perform this test to discover if the number of phases between the two groups of web services is significant in difference.

The mean number of phases mentioned by *data breach disclosed* web services was 4.5 with a standard deviation of 0.63. While the mean number of phases mentioned by *non-data breach disclosed* web services was 3.8 with a standard deviation of 1.0. Using a Mann-Whitney U test, we find a statistically significant difference in the mean number of phases covered by the two groups (U = 210, z = -2.217, $p = 0.027$). Using these test scores, we calculate an effect size $r = 0.294$, which is considered to be approximately a "medium" effect size [11, 20].

These findings may suggest that *data breach disclosed* web services have updated their account remediation advice once their compromised data was publicly known. The breach may have influenced a service to improve their systems and the resources they provide to users to secure their accounts. Interestingly, despite having the experience of a data breach, none of the web services which had disclosed a breach on haveibeenpwned explicitly mention reporting security flaws in the service to mitigate or prevent breaches.

Finally, we note that the analyses of differences of advice based on popularity and history of disclosing data breaches are preliminary and correlational. More work would be needed to confirm a causal relationship between a web service's coverage of account remediation advice and its popularity or history of disclosing data breaches.

5 Discussion

In this section, we discuss recommendations for implementing account remediation advice for web services. We also discuss what future work can be done to further this investigation.

Account Remediation Model: While remediation for each web service may have domain-specific concerns like fixing a playlist or recovering documents in cloud storage, our validated codebook provides evidence that the majority of account remediation steps are general, if not universal. Each phase in our codebook was constructed by analyzing multiple popular web services and creating codes that be broadly applied. We note that if one defines account remediation as "reversing the consequences of compromise," one must have all five phases for successful account remediation. One cannot claim an account is remediated until the compromise is discovered, user access is regained, the attacker has lost access, the account is restored to its pre-compromise state, and re-compromise is prevented. If any step is neglected, either a compromise is not remediated or the account will simply be re-compromised.

Our codebook also provides flexibility for domain-specific concerns as well. As discussed in Section 2, specific phases of account remediation such as discovering compromised accounts [6, 8, 19, 31, 33], recovering compromised accounts [3, 4, 16, 26, 32], and preventing compromises through general security practices [22, 30] have been researched and implemented. However, we are the first to conceptually define account remediation into a five-phase structured process. While the variations between services mean that account remediation advice cannot be totally centralized, we believe our codebook could be used for consumer advocates (such as the FTC) as the basis of public information campaigns and guides to help users in the complex task of account remediation.

On Service Responsibility: As mentioned in Section 4.2, much of the remediation advice given by services focus exclusively on account compromise resulting from other services or user error. They suggest that compromises may result from poor password choice, password reuse, falling victim to phishing, compromise of a "master" account like an email account, or malware infection. An example of advice following this tone is the following: "Don't worry, we have no indication that the Walmart systems have ever been compromised, but there are steps you should take to protect your personal information if you suspect unauthorized access or a phishing attempt". Services very rarely mention the possibility of a security flaw in their own service, even when they have previously disclosed a breach. While it may be the case that the source of most compromises is from external sources, companies should not completely shift responsibility onto individuals. Additionally, in some cases, users are limited in their ability to remediate an account. For example, banks do not allow users to unilaterally revoke a transaction after completion, and many web services automatically lock accounts based on indicators of compromise. An argument could be made that if

web services have the best visibility and ability to detect compromise, they should also be able to assist users proactively, if not automatically, in remediating the effects of that compromise. On the other hand, if it is true that account compromises mostly originate from external security problems, it would be unfair to put this burden solely on the web service. Similarly, the web service may have an incomplete perspective on what actions around the time of an account compromise were authorized or not. By analogy, credit card companies have regular monitoring for anomalous transactions, and in many cases can automatically block fraudulent transactions, even when caused by an external breach. Still, credit card companies often have to contact their users to confirm or deny specific anomalous charges. We recommend that web services consider to what extent they can automate remediating compromised accounts in order to balance responsibility with best serving users. We also suggest that language should be added to account remediation advice to encourage users to report security flaws with the service rather than focusing only on external causes of hacking.

Another question is what role, if any, law enforcement agencies have to play in identifying and prosecuting account compromises (especially in the furtherance of other criminal activities). We noted that 15 web services mention some form of evidence gathering of an account compromise alongside account remediation advice. However, we also note that computer crime is notoriously difficult to bring to prosecution, so it is arguable to what extent this would be helpful to current or even future victims.

Recommendations: Web services should, as a best practice, provide a mechanism to review account activity, including logins and actions that change the state of an account (purchases, password or preference changes, settings, user information.) Services should also provide better guidance on what "unusual activity" means through specific examples such as changed passwords, changed usernames, or changed emails. Owing to the large amount of prior work on account recovery, we recommend readers see the recommendations of prior work [3, 16, 26, 32]. All web services should also provide an interface to show all active log-in sessions and/or access permissions. This interface should also allow a user to revoke access for any or all current sessions. Along with the recommendation to show account activity, there should be an interface allowing users to revert changes made to their settings or remove unauthorized content. While not specifically coded for in our study, we observed that only six services provided a method to restore content deleted in an account compromise.

Enforcing mandatory customer service for account remediation purposes will inform the web service directly while also potentially discovering a large scale data breach. On the other hand, it potentially increases the effort on part of both the user and web service. Also, if mandatory customer service is not staffed 24/7, there may be consequential delays in preventing further damage from the compromise. This is why optional customer service may be a better feature to have, especially for complex remediation cases, because users without significant technical understanding of the compromise may need additional support. Finally, we observed a high variance in the prevention advice given by web services for what is largely the same problem, implying that many individual web services have incomplete prevention guidance. Similar to the work done by Redmiles et al. [29,30], there is an abundance of general prevention advice but a lack of advice prioritization.

Future Work: Future work should explore more usable or contextual guidance. Some of the steps in account remediation are technically complex to perform for users. Making the process of account remediation more usable and easier to follow will better aid users in remediating their accounts. For example, Facebook actually implements a chatbot-style wizard for guiding users through account remediation. It consists of easy to read diagrams that prompts users if they recognize information or settings on their account that is presented to them by the chatbot wizards. Future work could evaluate these approaches and explore ways of generalizing this approach to be usable for other types of web services beyond social media. Additionally, it is worth exploring to what extent a service could certify that an account has been remediated, or what assurances could be provided to users that their accounts have become "safe."

6 Conclusion

Online account compromises have become rampant, and anyone with an online account is susceptible to having their account compromised. The resources that help users remediate a compromised account should cover all the necessary procedures to help users re-secure their accounts. We investigated publicly available advice for account remediation from both top-ranked web services and lower-ranked web services. We identified important phases for account remediation that are not only sparse in coverage but also are not addressed by a significant amount of popular web services that provide account remediation advice. Also, the amount of web services we studied that even provide users with publicly available account remediation advice is critically low and did not surpass at least 15% of the total web services we analyzed that allow users to create accounts. Lastly, we discovered that highly ranked web services and web services with a previously disclosed data breach presented more complete coverage of their account remediation advice than other web services. Our analysis of the coverage of account remediation advice presented important areas that are lacking in attention, to which we explain credible recommendations to both bolster the advice and the process of account remediation.

References

[1] Devdatta Akhawe and Adrienne Porter Felt. Alice in warningland: A large-scale field study of browser security warning effectiveness. In *Presented as part of the 22nd USENIX Security Symposium*, pages 257–272, 2013.

[2] Rosaline S. Barbour. Checklists for improving rigour in qualitative research: a case of the tail wagging the dog? *British Medical Journal*, 322(7294):1115–1117, 2001.

[3] Joseph Bonneau, Elie Bursztein, Ilan Caron, Rob Jackson, and Mike Williamson. Secrets, lies, and account recovery: Lessons from the use of personal knowledge questions at google. In *Proceedings of the 24th International Conference on World Wide Web*, pages 141–150, 2015.

[4] Joseph Bonneau, Cormac Herley, Paul C Van Oorschot, and Frank Stajano. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *2012 IEEE Symposium on Security and Privacy*, 2012.

[5] Cristian Bravo-Lillo, Lorrie Faith Cranor, Julie Downs, and Saranga Komanduri. Bridging the gap in computer security warnings: A mental model approach. *IEEE Security & Privacy*, 9(2):18–26, 2010.

[6] Asaf Cidon, Lior Gavish, Itay Bleier, Nadia Korshun, Marco Schweighauser, and Alexey Tsitkin. High precision detection of business email compromise. In *28th USENIX Security Symposium*, pages 1291–1307, 2019.

[7] J Cohen. Statistical power analysis for the behavioural sciences. Hillsdale, NJ: Laurence Erlbaum Associates, 1988.

[8] Manuel Egele, Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. Towards detecting compromised accounts on social networks. *IEEE Transactions on Dependable and Secure Computing*, 14(4):447–460, 2015.

[9] Deen Freelon. *ReCal2: Reliability for 2 Coders*.

[10] Deen Freelon. *ReCal3: Reliability for 3+ Coders*.

[11] C Fritz. E Morris P, J Richler J. Effect Size Estimates: Current Use, Calculations, and Interpretation. *J Exp Psychol Gen*, 8:2–18, 2011.

[12] Kelsey R Fulton, Rebecca Gelles, Alexandra McKay, Yasmin Abdi, Richard Roberts, and Michelle L Mazurek. The effect of entertainment media on mental models of computer security. In *Fifteenth Symposium on Usable Privacy and Security ({SOUPS} 2019)*, 2019.

[13] Nethanel Gelernter, Senia Kalma, Bar Magnezi, and Hen Porcilan. The password reset MitM attack. In *2017 IEEE Symposium on Security and Privacy*, pages 251–267. IEEE, 2017.

[14] Mordechai Guri, Eyal Shemer, Dov Shirtz, and Yuval Elovici. Personal information leakage during password recovery of internet services. In *2016 European Intelligence and Security Informatics Conference (EISIC)*, pages 136–139. IEEE, 2016.

[15] Cormac Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *Proceedings of the 2009 Workshop on New Security Paradigms Workshop*, pages 133–144, 2009.

[16] Jun Ho Huh, Hyoungshick Kim, Swathi SVP Rayala, Rakesh B Bobba, and Konstantin Beznosov. I'm too busy to reset my linkedin password: On the effectiveness of password reset emails. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 387–391, 2017.

[17] Troy Hunt. *Pwned websites*. https://haveibeenpwned.com/PwnedWebsites.

[18] IBM. *IBM SPSS software*. https://www.ibm.com/analytics/spss-statistics-software.

[19] Hamid Karimi, Courtland VanDam, Liyang Ye, and Jiliang Tang. End-to-end compromised account detection. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 314–321. IEEE, 2018.

[20] Daniël Lakens. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in psychology*, 4:863, 2013.

[21] Megan Leonhardt. *The 5 biggest data hacks of 2019*, Dec. 17, 2019. https://www.cnbc.com/2019/12/17/the-5-biggest-data-hacks-of-2019.html.

[22] Peter Mayer, Yixin Zou, Florian Schaub, and Adam J Aviv. " Now I'm a bit angry:" Individuals' Awareness, Perception, and Responses to Data Breaches that Affected Them. In *30th USENIX Security Symposium*, 2021.

[23] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice. In *ACM on Human-Computer Interaction*, page 72, 2019.

[24] Jeremiah Onaolapo, Enrico Mariconti, and Gianluca Stringhini. What happens after you are pwnd: Understanding the use of leaked webmail credentials in the

wild. In *Proceedings of the 2016 Internet Measurement Conference*, pages 65–79, 2016.

[25] Bijeeta Pal, Tal Daniel, Rahul Chatterjee, and Thomas Ristenpart. Beyond credential stuffing: Password similarity models using neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 417–434. IEEE, 2019.

[26] Simon Parkin, Samy Driss, Kat Krol, and M Angela Sasse. Assessing the user experience of password reset policies in a university. In *International Conference on Passwords*, pages 21–38. Springer, 2015.

[27] Peng Peng, Chao Xu, Luke Quinn, Hang Hu, Bimal Viswanath, and Gang Wang. What happens after you leak your password: Understanding credential sharing on phishing sites. In *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security*, pages 181–192, 2019.

[28] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. Tranco: A research-oriented top sites ranking hardened against manipulation. *arXiv preprint arXiv:1806.01156*, 2018.

[29] Elissa M Redmiles, Amelia R Malone, and Michelle L Mazurek. I think they're trying to tell me something: Advice sources and selection for digital security. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 272–288. IEEE, 2016.

[30] Elissa M Redmiles, Noel Warford, Amritha Jayanti, Aravind Koneru, Sean Kross, Miraida Morales, Rock Stevens, and Michelle L Mazurek. A comprehensive quality evaluation of security and privacy advice on the web. In *29th USENIX Security Symposium*, pages 89–108, 2020.

[31] Xin Ruan, Zhenyu Wu, Haining Wang, and Sushil Jajodia. Profiling online social behaviors for compromised account detection. *IEEE transactions on information forensics and security*, 11(1):176–187, 2015.

[32] Stuart Schechter, AJ Bernheim Brush, and Serge Egelman. It's no secret. measuring the security and reliability of authentication via "secret" questions. In *30th IEEE Symposium on Security and Privacy*, pages 375–390. IEEE, 2009.

[33] Richard Shay, Iulia Ion, Robert W Reeder, and Sunny Consolvo. " My religious aunt asked why I was trying to sell her viagra" experiences with account hijacking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2657–2666, 2014.

[34] Kurt Thomas, Frank Li, Ali Zand, Jacob Barrett, Juri Ranieri, Luca Invernizzi, Yarik Markov, Oxana Comanescu, Vijay Eranti, Angelika Moscicki, et al. Data breaches, phishing, or malware? Understanding the risks of stolen credentials. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1421–1434, 2017.

[35] Kurt Thomas, Jennifer Pullman, Kevin Yeo, Ananth Raghunathan, Patrick Gage Kelley, Luca Invernizzi, Borbala Benko, Tadek Pietraszek, Sarvar Patel, Dan Boneh, et al. Protecting accounts from credential stuffing with password breach alerting. In *Proceedings of the 28th USENIX Security Symposium*, pages 1556–1571, 2019.

[36] Courtland VanDam, Jiliang Tang, and Pang-Ning Tan. Understanding compromised accounts on twitter. In *Proceedings of the International Conference on Web Intelligence*, pages 737–744, 2017.

[37] Ke Coby Wang and Michael K Reiter. Detecting stuffing of a user's credentials at her own accounts. In *29th USENIX Security Symposium*, pages 2201–2218, 2020.

7 Codebook

Compromise Discovery		
Codes	**Code Explanations**	**Examples**
Billing/finance issues	Unwanted changes in financial or billing settings/standings or unauthorized credit card charges.	You see charges or notices for purchases that you didn't make.
Email changed	Observe any email associated with account has been changed.	What do I do if someone changed my email address?
Explicit notification	Service notifies you of login or possible compromise by email or other factor. Check this if the service sends emails about new logins.	You receive an email or notification that your Apple ID was used to sign in to a device you don't recognize or did not sign in to recently (for example, "Your Apple ID was used to sign in to iCloud on a Windows PC").
Account locked by provider	Cannot access account due to account being locked or disabled.	For your protection, we may place a temporary hold on your account.
Account otherwise unavailable	Account is not accessible due to circumstances outside of provider locking account.	You can't sign in for another reason.
Observed unauthorized logins	Includes if "observation" is due to a notification from the service, but not exclusively.	You see logins from unexpected locations on your recent activity page.
Password changed	Observe password associated with account has been changed.	Someone changed the password on my Etsy account.
Social media or third party account connected	Unwanted social media becomes associated with account.	A malicious application has been given access to your account.
Unauthorized/supicious activity	Including changed content on streaming sites, but must be more than login. For example messages, friend requests, playlists, etc.	If you notice unfamiliar activity on your Google Account, someone else might be using it without your permission. Use the info below to help spot suspicious activity.

Account Recovery		
Codes	**Code Explanations**	**Examples**
Customer service process	Engage with service customer support (chat client, form, email, etc) to regain access/reset password.	If you can't access your account and believe that someone else has accessed it, complete the form and after receiving it we'll verify that it's your account and then help you regain access.
Password reset	Initiate a password reset challenge or go through password change process.	Change your password immediately.
Run endpoint security	Run external security applications on computer to stop a suspected *ongoing* attack.	If you see any successful sign-in that you do not recognize, run a scan with your security software and remove any malware you find.

Limiting Access		
Codes	**Code Explanations**	**Examples**
Remove third party access	Disallow external third party applications (including social media) from accessing account.	Revoke access to any suspicious third-party apps.
Review active session	Review activity/logs for currently active sessions to see if compromise is ongoing.	Review your active sessions to see all the places you're signed into LinkedIn right now.
Sign out everywhere (specific function)	Logs out *all* instances of account (not just one or a few).	We recommend to log out of all computers from your phone.
Sign out of unknown session	Logs out of individual unrecognized instances of account.	If your account does get hacked, you can remove any trusted devices that you didn't log in to yourself.

Service Restoration		
Codes	**Code Explanations**	**Examples**
Customer service process	Engage with service customer support (chat client, form, email, etc.) to help restore data etc.	Contact us for help removing unauthorized bids or listings.
Fix logs of past viewing/activity/content history	For example, viewing history, input to recommendations, past purchases.	Review Order history for unrecognized charges.
Review and/or remove activities/content	For example, deleting friends you didn't add, messages you didn't write.	Delete any resources on your account that you didn't create, such as EC2 instances and AMIs, EBS volumes and snapshots, and IAM users.
Verify settings	User should verify security, privacy, or account settings.	Review your general account settings to make sure all other information is correct.
Verify user information	User should check the identifying information for users (email, name, address, or payment info like credit card number).	Verify that the email address and mobile number associated with your account are accurate in Snapchat settings.

Prevention		
Codes	**Code Explanations**	**Examples**
Advice about secure email	Describes advice on suspicious emails, phishing, etc.	Phishing is when someone tries to trick you into giving up your Twitter username, email address or phone number and password, usually so they can send out spam from your account.
Always log out on shared devices	Always log out shared instances of account.	Sign out of public computers- - -Always sign out of your accounts when you're done.
Check/modify related accounts	For example, email accounts, shared passwords, etc.	Check your personal email account(s) tied to your account to ensure their security.
Enable 2FA	Enable any 2FA for every login attempt.	Enable Two-Factor Authentication (2FA).
Enable phone-based recovery	Enable ability to *recover* account/credentials by using a phone number as a second factor.	Add a recovery phone number to your account so that you can get back into your account faster and keep your account more secure.
Keep software up to date	Catchall: any application/program/devices/software up to date with current updates.	Regularly patch, update, and secure the operating system and applications on your instance.
Password advice: strong, unique, change frequently	Catchall for any password advice (good bad or otherwise).	Create a strong password. Make it unique: Do not reuse an existing password when setting up an account for PlayStation Network.
Physical security	Catchall for any advice to maintain physical security of devices, environment, etc.	Don't leave your devices unlocked or unattended where anyone can use it.
Remove access to third party apps	Prompted to disallow external third party applications from accessing account.	Remove suspicious applications or browser add-ons.
Run endpoint security solutions	Run external security programs/applications on computer to prevent *future* attacks.	Always use an antivirus program to check the files you receive from other people.
Sign out of devices	Log out of *individual* devices that have instances of account.	Log out when you are done.

8 Web Services Studied

Very Popular Websites		Less Popular Websites	
Ranking	Website	Ranking	Website
1	google.com	524	hootsuite.com
2	facebook.com	542	ox.ac.uk
3	youtube.com	547	umn.edu
4	microsoft.com	559	uci.edu
5	twitter.com	568	ucla.edu
7	instagram.com	575	att.com
9	netflix.com	578	snapchat.com
10	linkedin.com	608	uchicago.edu
13	wikipedia.org	620	playstation.com
14	apple.com	635	xfinity.com
18	yahoo.com	658	parallels.com
23	pinterest.com	669	epicgames.com
25	vimeo.com	682	fidelity.com
28	reddit.com	730	ning.com
40	amazonaws.com	776	verizon.com
44	tumblr.com	785	uber.com
45	godaddy.com	795	msu.edu
51	skype.com	806	ea.com
55	whatsapp.com	836	northwestern.edu
56	dropbox.com	837	crunchyroll.com
58	soundcloud.com	886	arizona.edu
61	myshopify.com	904	wattpad.com
67	twitch.tv	917	stripe.com
79	spotify.com	932	namecheap.com
81	paypal.com	942	xbox.com
93	cloudflare.com		
94	ebay.com		
117	etsy.com		
170	aol.com		
183	fandom.com		
188	walmart.com		
209	yelp.com		

9 Data

Our annotated advice is available at: https://github.ncsu.edu/lcneil/Investigating-Web-Service-Account-Remediation-Advice

Knowledge and Capabilities that Non-Expert Users Bring to Phishing Detection

Rick Wash
Michigan State University

Norbert Nthala
Michigan State University

Emilee Rader
Michigan State University

Abstract

Phishing emails are scam communications that pretend to be something they are not in order to get people to take actions they otherwise would not. We surveyed a demographically matched sample of 297 people from across the United States and asked them to share their descriptions of a specific experience with a phishing email. Analyzing these experiences, we found that email users' experiences detecting phishing messages have many properties in common with how IT experts identify phishing. We also found that email users bring unique knowledge and valuable capabilities to this identification process that neither technical controls nor IT experts have. We suggest that targeting training toward how to use this uniqueness is likely to improve phishing prevention.

1 Introduction

Email is one of the most commonly used methods of communication, especially in large organizations and for e-commerce. Over 3.9 billion people have email accounts, and collectively they send and receive over 290 billion emails per day [11]. Email is one of the major methods that is used to communicate with strangers. However, because email is a global system where anyone can communicate with anyone, malicious actors send emails that pretend to be something that they are not, and trick people into taking actions that they otherwise wouldn't — which is known as phishing [34].

Phishing messages are an attack vector that has caused a large amount of damage in society. Phishing emails have been used to steal large amounts of money [22], install ransomware [31], or simply steal email contents that are later made public [21]. 32% of all corporate breaches in 2018 were due to phishing [33]. Spear-phishing – a variant where emails are custom targeted to the recipients – is used by 65% of groups doing targeted cyber-attacks, and is more commonly used than zero-day vulnerabilities (only 23% of such groups) [32].

Phishing is a socio-technical problem, and addressing the problem requires the coordinated work of both technological innovation and human intervention. Technologies are being developed that help identify and filter phishing messages, but these technologies do not work with 100% accuracy and can be slow to respond to new innovations by adversaries [14]. IT administrators and governments often try to stop phishing before it starts by disrupting phishing websites and bulk email sending [10]. But the last line of defense is the end user; phishing messages that go through these other defenses can still be detected or ignored by end users to prevent harm.

In this paper, we surveyed email end users without IT training or expertise and asked them about specific experiences with phishing emails they have received. Approximately half of survey respondents were able to identify a specific incident that they then answered detailed questions about. Building on Wash's [34] model of how IT experts detect phishing emails, we asked each person about what they noticed about the email, what they expected in the email, what made them suspicious of the email, what investigation they did, how they decided whether the email was legitimate, and what they finally did with the email.

From these questions, we are able to identify patterns in how email users who are not IT experts currently identify phishing scam emails in their own inboxes. Most research looks at phishing detection failures and what needs to be fixed; instead we compare non-experts with Wash's experts and identify what is working well that we can build upon. We find that email users often bring unique knowledge to this identification process that other phishing prevention methods do not have, such as whether the email was expected or not

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

and what emails like this typically look like and ask for. We also find that email users have valuable capabilities for investigation, such as asking other people for advice, or checking with senders for validity. Together, these findings suggest that email users can be an important part of the phishing prevention ecosystem, though phishing training can be improved to focus on how users can better use their unique knowledge and capabilities.

2 Previous Work

2.1 Preventing Harm from Phishing

Our society has three forms of defenses that help identify and limit the success of phishing scams. Technological defenses try to automatically detect known features of phishing emails and block or remove emails. Some defenses combine the work of computers and people by warning end users of the potential phishing message, which is then investiagted further by the end user to determine if it is a phishing email. And finally, there are human defenses, where the recipient of the email is relied upon to recognize the email as dangerous and act accordingly.

2.1.1 Automated Detection and Deletion

Automated detection and deletion approaches aim to classify emails as phishing or legitimate and block or remove them before the end user encounters them. Efforts in this space have focused on improving and finding new ways to identify outgoing and incoming phishing messages using blacklists [10], heuristics [3, 13, 16, 23], and machine learning [9, 29]. These approaches filter emails based on known features that conclusively identify emails as phishing.

Automated approches, however, rely on probabilistic algorithms which produce false positives, causing legitimate emails to be blocked or removed. In addition, automated approaches have limited ability to detect new permutations of phishing attacks [12] and cannot identify all older phishing emails.

2.1.2 Phishing Warnings

Phishing warnings augment automated detection techniques by warning end users of potential phishing emails, instead of blocking or removing them. Warnings are commonly used when automated detection cannot conclusively classify an email as phishing [25]. In practice, warnings have been reported to improve end users' ability to identify phishing emails [8, 26]. Ongoing research efforts in this area have focused on finding better ways to design and present warnings to the end user.

Despite their positive impact, warnings share the same limitations with automated detection and deletion approaches. They are prone to false positives (tagging legitimate emails as potentially dangerous) and false negatives (letting malicious emails through without warning, especially zero-hour phishing attacks). As Yang et al. argue, warnings and user training must complement each other to improve their effectiveness [37].

2.1.3 User Training

Security researchers and practitioners have developed various methods and materials for training users to identify and react to phishing emails accordingly. Kumaraguru et al. [19] and Caputo et al. [2] found that embedded training (i.e. instructional materials presented the moment a participant clicked on a URL in a phishing email), which is very commonly used in large organizations, improved user motivation to learn and enhanced knowledge acquisition. Rader et al. [27] found that people also learn about phishing scams and protective actions from stories about security incidents. Wash and Cooper [35] found that traditional facts-and-advice phishing training worked better when presented by an expert, while narrative security stories worked better when told by a peer.

The most widely shared phishing training messages across governments, businesses, and individuals teach people to identify certain cues (e.g. sender email address, URLs in emails, poor grammar or spelling) or apply a set of rules to detect, avoid and report phishing messages. Such training messages have been extensively studied and have shown potential to improve people's resistance to phishing attacks [4, 19]. Some messages focus on behavioral change, e.g., never click on a URL or open an attachment in an email from an unknown sender.

Other training messages focus on informing users of the common types of phishing threats and how to identify them, with the aim of manipulating the risk level and subsequently the level of fear in the users [5, 20]. Some researchers have argued that fear appeals increase end users' intentions to act securely. However, despite their ability to change behavioral intentions of end users [5], fear appeals do not predict or result in secure behavior [6].

User training typically focuses on aspects of the email message and tries to change the way people think about email messages so that they are paying attention to the features most associated with phishing. Studies have shown that this improves user knowledge, enhances their capabilities to identify phishing emails, and reduces the number of successful attacks [2, 19, 35]. However, the number of successful phishing attacks is still reasonably high, comprising 32% of all corporate breaches in 2018. More needs to be done to improve the capabilities of end users in identifying and preventing phishing attacks.

Most user training is developed from understanding how and why people fall for phishing [6]. We postulate that if training were to focus more on aspects of how people already

think about and deal with email in general, this can open up new avenues for phishing training. Unfortunately, we do not have a comprehensive understanding of how non-expert users do this. A similar problem was encountered in technical skills training where researchers investigated ways to improve the training of troubleshooters (technicians) [15]. They studied and identified a common conceptual process and strategies that technicians used when troubleshooting problems. This helped them to identify gaps in existing training methods and messages and subsequently helped them to identify areas of improvement. We argue that undertstanding the process(es) and strategies that non-experts use to identify phishing emails can reveal potential improvement areas for phishing training.

2.2 How Do People Identify Phishing Emails?

Downs et al. [7] investigated decision strategies of non-expert computer users when encountering suspicious emails. They identified three strategies that participants used to make sense of the emails they received: 1) this email appears to be for me; 2) it's normal to hear from companies you do business with and 3) reputable companies will send emails. Downs et al. [7] state that none of the strategies helped people to identify well-constructed phishing messages. The study, however, involved role-playing in a controlled environment. We do not know which of these strategies apply to and how prevalent they are in people's natural contexts and inboxes.

Wash [34] looked at how experts identify phishing emails by interviwing 21 IT experts about instances when they successfully identified emails as phishing in their inboxes. He identified a 3-stage process for identifying phishing emails. In the first stage, the email is received and treated like any other email — the content in the email is taken at face value and the person tries to make sense of the email and figure out what it is asking them to do. As they do this, they notice discrepancies — things that "feel off" about the email. Eventually, something triggers the person to think that this email is not legitimate — that it might be a phishing email that is not what it says it is. At this point, they become suspicious and begin explicitly looking for things that can help them determine if the email is legitimate or not. These new pieces of information often allow them to conclusively identify the email as phishing.

The work of Wash [34] demonstrates how some of the lessons from phishing training are applied in real-world contexts. However, Wash studied experts only. Experts might have more advanced skills, experience and knowledge about phishing and countermeasures compared to non-experts. We do not know which of the findings might apply to non-experts and can be used to improve their training.

2.3 Phishing: A Socio-Technical Problem

Phishing is a socio-technical problem. Automated solutions do not detect 100% of phishing emails. Hence end users must identify these emails in their inboxes. As Khonji et al. state, no single solution exists to mitigate phishing attacks [17]; thus automated / warning and user training techniques must be implemented to complement each other [19]. This is comparable to James Reason's Swiss Cheese Model (SCM) [28] of accident causation and response. SCM is a popular tool used to investigate or analyze the complexity of systems by showing that an incident is a result of a combination of active failures by operators and latent conditions of the system. SCM depicts socio-technical systems as multiple slices of Swiss cheese that are stacked together, each slice with a hole. Each slice depicts a layer of system defense against certain types of failures, while each hole represents failures in system defense at that particular layer. Bryans and Arief applied the model to understand security layers and fault-tolerance in computer systems [1]. They depict each layer as a protective mechanism against certain types of attacks, but has weaknesses (holes) against other types.

Both automated detection and deletion and warning techniques rely on the end user as the last line of defense against phishing. However, the number of recent successful phishing attacks suggests that more work needs to be done to improve user training. While most training focuses on teaching end users to identify known, conclusive features of phishing emails, Downs et al. [7] and Wash [34] found that end users rely on features other than conclusive distinguishers to identify phishing emails. We need to explore improved ways of keeping the user in the loop of defending against phishing attacks. More research needs to be done to understand how non-experts identify phishing emails, what aspects or information they rely on, and the kinds of things they do in the process. This understanding can help us to tailor and target phishing training and technologies that support human decision-making. Our study takes a first step in this direction by applying Wash's model in a survey to study the techniques that non-experts follow to identify phishing emails.

3 Methods and Sample

In this paper, we look at how non-expert users identify phishing emails, and look at whether some of the techniques that Wash [34] identified in experts continue to be present when non-experts identify phishing emails. To study this, we conducted a survey where we asked non-expert Internet users to remember a specific email that they received that was "suspicious or potentially harmful," and then answer questions about their experience with that email.

We asked questions to try to understand what they noticed and didn't notice about the emails respondents received and understand what kinds of things seemed important to them.

This is a retrospective account of a past email; we expect that respondents won't remember some of the details of what happened. We make the assumption that things they don't remember are most likely less important in their thinking about the email [18].

3.1 Survey

We started with a survey instrument that is loosely based on Rader et al. [27]. Near the beginning of the survey, we asked respondents to identify a specific "story" or incident where they received a suspicious or potentially dangerous email. We then asked them to answer a number of questions about that specific incident.

We included a screening question that asked potential respondents whether they could recall receiving the type of email we were interested in. The survey informed respondents that "In this survey, we are interested in hearing about emails you received that were suspicious or potentially harmful in some way." It then asked them to think back over their email, and told them it was OK to look back at their email if it would help. We asked "Can you remember any suspicious or potentially harmful email messages that you've received?" Only respondents who answered yes to this question proceeded on with the survey. 315 potential respondents that were otherwise qualified were excluded from the study because they did not answer "Yes" to this question.

Much like Rader et al. [27], we began the survey with an elicitation process to get respondents to identify a single "suspicious or potentially dangerous email" to answer questions about. The elicitation included three parts. First we asked respondents to write down in a short answer box "ways that an email message can be unsafe or cause security problems" and "ways you know of to recognize an email that is suspicious or potentially harmful." These prompts were intended to help trigger the respondent's memory of potential phishing emails. Respondents wrote an average of 12-14 words for each of these prompts.

Second, we asked the respondent to "think about times in the past when you personally received a suspicious or potentially harmful email" and "list as many of these emails as you can remember" in a text box. Respondents averaged 15 words in response to this prompt.

Third, we presented this list back to the respondent and asked the respondent to "Choose one email message from the list above that it's easy for you to recall details about." We asked them to briefly summarize that specific email. We presented this brief summary back to the respondent at the top of each subsequent page of the survey to help them remember which email they were answering questions about. These summaries averaged 21 words long.

The rest of the survey asked for more details about the specific email incident that was chosen by the respondents. Based on Wash's model [34], we identified six processes that experts use in phishing detection. We structured the questions around these six processes:

- **Noticing**: Things they noticed about the email, like when they received the email, what kind of mail (attachments, etc.), work or personal content, work or personal account, etc.
- **Expecting**: What they were expecting in the email; builds on noticing and compares what they noticed with what they expected. Have they received other emails like this, interacted with sender before, was the email expected, etc.
- **Suspecting**: What felt "off" about the email — subject, from, body, etc.. What in the email caused them to suspect the email. Did it contain links, attachments, etc.
- **Investigating**: What they went and explicitly looked for once they suspected the email (if anything) to figure out if the email was legit or fraud. Things like "did you look at headers, or hover over links, or try to contact the sender?"
- **Deciding**: How was the legit/phish decision made. Did you decide, and if so, how? How sure are you?
- **Acting**: After deciding, what did you do with the email? Report it? Just delete it? How did you feel about the email? Fear? Dread? Anxiety?

The complete survey instrument can be found in the supplementary materials.

3.2 Sample

We contracted with Qualtrics to field our survey to a panel of US participants in February 2020, which was just before the COVID pandemic. We excluded respondents who had had technical expertise or worked as technology professionals because we specifically wanted non-expert respondents. We placed quotas on age, gender, and ethnicity that roughly matched the US population, to try to get a more representative sample. We received a total of 297 valid responses. Respondents were compensated by Qualtrics with points that could be redeemed for items.

Table 1 summarizes the demographics of our sample. Our sample achieved the quotas and therefore roughly matches the US population along those lines. It also happened to come close to the US population in terms of education.

Only about 50% of our sample was currently employed either full-time or part-time. This is lower than in the US population (which was approximately 61% employed at the time of the survey [24]). This is the major way we believe our sample differs from the larger US population. We are not sure how this might affect responses about phishing emails.

	N	%		N	%
Age			**Employment**		
18-30	75	25%	Employed Full Time	105	35%
30-50	104	35%	Employed Part Time	42	14%
50-65	73	25%	Unemployed and looking for work	24	8%
Over 65	45	15%	Unemployed and not looking	25	8%
Gender			Retired	45	19%
Man	151	49%	Disabled	29	10%
Woman	156	50%	Student	16	5%
Other	2	1%	**Annual Household Income (USD)**		
Prefer not to answer	1	0%	Less than $25,000	66	22%
Ethnicity			$25,000 to $34,999	51	17%
White	202	64%	$35,000 to $49,999	35	12%
Hispanic, Latino, or Spanish	51	16%	$50,000 to $74,999	69	23%
Black or African American	37	12%	$75,000 to $99,999	33	11%
Asian	18	6%	$100,000 to $149,999	30	10%
American Indian or Alaska Native	8	3%	$150,000 to $199,999	7	2%
Education			$200,000 or more	6	2%
No College	71	24%			
Technical, Trade, or Vocational	22	7%			
Some college	102	34%			
College Degree	102	34%			

Table 1: Demographics of the survey sample. We received valid responses from a total of 297 respondents. Quotas were used on Age, Gender, and Ethnicity to approximately match demographics of the United States.

The majority of respondents in our sample had previous experience with cybersecurity incidents; only 17% of respondents indicated that they had not been a victim of a cybersecurity incident. About half of the sample reported having a virus (52%), and almost half reported having received a notification of a data breach (47%). Approximately one quarter (26%) had been the victim of credit card fraud, and 6% reported being a victim of identity theft more serious than credit card fraud. 18% reported having a device hacked. Interestingly, 16% of respondents reported having previously fallen for a phishing email or other scam email. These statistics suggest that our sample is also somewhat biased toward people who have had prior experience with cybersecurity incidents.

3.3 Analysis

Near the end of the survey, we asked respondents to "please write the story of the email as if you were telling it to a friend." We provided a large text box for the participant to enter in the story, and required that respondents enter at least 300 characters into this box. Respondents averaged over 400 characters (mean=411, min=300, max=1523), which is about 80 words per story on average (mean=81, min=41, max=288). We had two research assistants code these stories in parallel, meeting weekly to update the codebook, measure agreement, and resolve differences. We ended up with a codebook that coded stories for features organized in 5 categories: properties of the purported sender of the email; the action requested by the email; what felt off in the email; actions taken in the story; and final decision about the email.

After the training and codebook development, the two coders coded all 297 stories independently for a codebook of 39 distinct codes. After this initial coding, over half of the codes had a Cronbach's alpha above 0.7, and only 3 codes had an alpha below 0.5. We dropped the 3 codes with low agreement. The two coders then met and talked through all instances where there was disagreement and mutually agreed to a final decision about all codes for all stories.

In this paper, results from this manual coding will be explicitly labeled as such. Any results not labeled as resulting from manual coding are self-report data directly from questions in the main body of the survey. 13 (4%) of the stories were agreed to be "not a story" by both coders. These were instances where the participant filled out this text box for the whole survey, but did not describe an experience with a specific email, and instead described more general experiences. These responses are not included in statistics for the manual coding.

Replication materials for this analysis are available at https://osf.io/82sd9/. Additionally, all stories are presented exactly as they were entered by respondents, typos included.

4 Findings

In this survey, we asked respondents to identify "a suspicious or potentially harmful email message you received in the past." 315 otherwise qualified respondents were unable to identify an email, and 311 otherwise qualified respondents were able to do so. Quotas only applied to the qualified respondents who remembered such emails, and respondents were incentivized to remember such an email to participate in the survey and receive the incentive payment. Our goal was not to discover how prevalent phishing is among different demographic groups, and this sample should not be interpreted as measuring prevalence of phishing. However, it suggests that approximately 50% of the non-expert people in the Qualtrics subject pool have stories about specific phishing emails that they have received, which shows how widespread experience with these emails is.

Almost all of the remaining questions on the survey then asked the respondent for more details about the specific incident where they received that email that they chose to tell us about: what happened as they received it, what did they notice, and how did they handle it? In the majority of this paper, we report statistics about responses to multiple choice questions.

Based on findings from Wash [34], we organized the survey based on six different activities that a person needs to do to recognize a phishing email: 1) Noticing aspects of the email; 2) Forming expectations about what should and should not be in the email; 3) Becoming suspicious of the email; 4) Investigating the email; 5) Deciding whether the email is suspicious or not; and 6) Acting on that decision.

These six activities provide a way for us to describe what generally happens when a person receives a phishing email, and to look at patterns in what they notice and what they do. We organize our description of the findings in this paper around these six different activities.

4.1 Incidents

Each participant was asked to answer questions about a single incident that they experienced. We begin by describing the types of incidents that respondents reported on. Each incident was an email that the participant had received and decided was suspicious or potentially dangerous. All of these incidents represent emails that had made it through any technical defenses and into the participant's inbox, and so do not include phishing mails that were successfully filtered by technical phishing protections. Still, these emails were not uniform; respondents reported receiving a wide variety of different types of phishing scam emails.

We asked each respondent to identify a list of possible incidents / emails that would qualify, and then asked them to choose one that is "easy for you to recall details about" and then answer more questions about that one. We had a total of five questions that tried to understand broadly what these emails were about — one question near the beginning asking the respondent to summarize the incident, one question near the end asking the respondent to explain the whole incident, and then three questions asking for brief, 5-words descriptions of the chosen incident. Here we use these 5-word descriptions to describe the kinds of incidents that people reported on.

When asked to summarize the incident early in the survey, respondents responded with an average of 21 words (median: 17 words). In these summaries, respondents mostly reported facts about the email that they received, with the most common words being email (39% of respondents), account (17%), money (15%), link (13%) and received (11%).

In addition to the summary, we asked respondents, "In approximately five words" to describe what made the email suspicious, what made the email hard to figure out, and what the email was asking them to do. The respondents reported that they were suspicious mostly looking at the email / sender address or because it involved money. The emails were mostly asking respondents to click links (22%), for money (17%), or for "information" (14%). Together, these summaries suggest that most of the phishing stories were about economic issues (money) or asking for or providing information.

81% of respondents indicated that they found it easy to remember such an email. The emails that respondents chose to respond about were widely distributed in time: 24% of respondents received it within the last week; 30% within the last month (but not the last week); 25% within the last year (but not last month); and 15% more than a year ago.

In the manual coding, we coded the full incident stories for information about who the purported sender of the email was. This was not who actually sent the email, but who the email pretended to be from. 44% indicated that the email was from a group or organization, and 25% indicated that the email seemed to be from an individual. In 30% of the stories, the participant indicated that they had a pre-existing relationship with the purported sender, and 14% of the stories the participant explicitly stated that they did not have a pre-existing relationship. 76% of the pre-existing relationships were with a group or organization; suggesting that emails pretending to be from an organization were more likely to be seen as part of a pre-existing relationship.

As an example of a story about an email from an organization the participant had a pre-existing relationship with, consider the following story about an email from Amazon.com:

P233 Story: *I received an email that appeared to be from amazon. It had my name and address but said i owed money for a purchase. I hadn't purchased anything for a while so that seemed strange. Email had misspellings and an odd looking link. I looked closely at the email, then checked my amazon account on their website. There was nothing there about any orders or owing money.*

The actual senders varied widely across stories: about 12%

said it was a bank or financial institution, 8% said the email appeared to be from a foreign person, 4% from the government, and 2% from an IT support organization.

In the manual coding of stories, we also coded for what kind of information was being requested. 30% of the stories mentioned that the recipient of the email would receive some sort of valuable (money, award, gift, job offer, etc.), and 19% of the stories reported that the email asked the recipient to send money. 19% of the stories mentioned that the email was asking for personal information, 10% of the stories were asking for technical information such as usernames or passwords, and 10% of the stories were asking for financial information like bank account numbers, credit card numbers, etc. This suggests our respondents received emails with a wide range of requests, with no particular type of request being overwhelmingly common. What end users consider to be phishing is diverse, and training that focuses mostly on cues may miss classes of email messages that stand out to end users as potentially harmful.

4.2 Noticing

4.2.1 What people notice in an email

As a person reads an email, they cannot notice and remember everything about the email. Instead, the things in the email that the person can most easily make sense of and connect with are the easiest to notice and remember [18]. We asked respondents "What aspects of the email stood out to you?" and allowed them to check all that apply. The answers to this question show us, for these suspected phishing emails, what aspects of the email were most important to the respondents, because they were the most memorable.

By far, the aspect noticed by the largest number of people was that the email included a request for an action. 76% of respondents noticed this about the email. This corresponds well with past research that suggests that people tend to use email as a to-do list [36]; they quickly focus on what the email is asking them to do. It also corresponds with Wash's [34] finding that requests for actions (action links) were important triggers for experts.

The second most commonly noticed aspect of email was what the email was about, with 52% of respondents noticing this. The topic of the email, and whether that topic is relevant to the recipient of the email, is commonly seen as an important aspect of phishing. This data backs up that idea, and shows that this is something that people quickly are able to identify and remember about emails.

Much past work on phishing has focused on "conclusive distinguishers": aspects of an email that can help the recipient to conclusively distinguish legitimate emails from phishing emails, or at least strongly indicate phishing. For example, phishing training usually focuses on aspects such as inappropriate URLs in links, urgency in requests for action, or poor grammar/spelling. However, Wash emphasizes that when experts identify phishing emails in their own inboxes, they instead look for more minor discrepancies, which are things that seem off about the email, but don't necessarily indicate phishing and definitely are not enough on their own to conclusively identify phishing.

These first two things that respondents noticed — requests for action and topic of the email — do not conclusively indicate that the email is a phishing message, and are not normally part of phishing training. Instead, they simply indicate that there is something weird about the emails. However, for some people they might be enough. For example, consider this story:

P19 Story: *I got an email last Friday from one of the companies we work for that pays us to provide service for them and I immediately could tell it was a fake email because the company the email sender disguised themselves as is a company that pays us, we don't pay them.*

I called the company we work for and reported it to them so they would know someone was trying to disguise themselves as them

The next two most commonly noticed aspects of the email are much more commonly associated with phishing identification: links in the email (44%), mistakes or poor quality (41%). These are often found in phishing emails (especially the kinds of phishing emails that non-experts in our sample might be able to successfully detect).

38% of respondents reported that the sender's name stood out to them. The remaining aspects of email, such as attachments, images, formatting, or length of email, were noticed by less than 20% of respondents, though all of them were important to a non-trivial subset of users. This finding suggests that people seem to naturally notice actions and topics of email much more than they notice more conclusive distinguishers like URLs or typos. This is important, because a person cannot use a feature to detect phishing unless they first notice that feature.

4.2.2 Non-email features

In addition to noticing aspects of the email, there are a number of aspects of the situation that are not necessarily part of the email but nonetheless appear to be important and memorable to respondents.

90% of the respondents noticed that the email had come to their personal email account. None of our respondents chose the "I don't remember" option for which email account it arrived at. The account that the email arrived to is salient and memorable to respondents, and is possibly something that can be used to help identify suspicious email. Only 78% of respondents reported that the email was of a personal nature.

70% of the respondents reported that the email appeared to come from a company, business, or other organization(i.e.

from a person). Only 6% of respondents cannot remember who the email appeared to come from. The email sender appears to be a highly salient aspect of the email. It is interesting that 94% of respondents can remember who the email appeared to come from, but that fact only stood out to only 38% of them.

4.3 Expecting

When trying to understand and make sense of an email, people naturally fall back to what kinds of email they expect to receive, and to comparing the email with past emails that they have received [34].

Almost all of the suspicious emails arrived unexpectedly (95%). This seems to be one of the strongest aspects of phishing identification for our respondents. It is also something that users find relatively easy to identify, but is almost impossible to measure technically. That is, whether an email is expected or not is something that is a valuable piece of information that only the user has and computers do not.

However, just because the email was unexpected does not mean it was unfamiliar. 72% of respondents reported either "somewhat agree" or "strongly agree" to the statement "I felt like I had received other email messages like this one before.". That is, almost three quarters of the emails felt familiar to the recipients.

This fact both helps and hinders phishing detection. On the one hand, since the emails are familiar, people can easily integrate these into their lives and might not read them very carefully. On the other hand, as Wash [34] points out, when the email is similar to other, past emails, then it is possible to form expectations about what is typical in those past emails, and then compare this email to the past, similar emails and notice more things that are different or wrong about this email.

While respondents reported receiving emails similar to the suspicious email, the suspicious email was not a typical email. 86% of respondents chose "somewhat agree" or "strongly agree" about the statement "This email message seemed different from the email messages I typically receive."

Putting these findings together, suspicious emails that people remember are generally emails that are unexpected, different than the emails typically received, but often are like other emails that have been received before. The feeling that an email is suspicious, or unexpected, represents intuition, or a "gut feeling" about an email, and such intuitions are often important aspects of human decision-making [18].

Only 19% of respondents remembered receiving an email from this sender before. The remaining either had never received an email from the sender (45%) or were not sure (33%). So while the email felt familiar, the sender generally was not. Even more telling, only 12% of respondents had actually interacted with the sender before reading this email, and 80% of respondents checked "No" to having previously interacted with the sender. This suggests that non-experts remember and pay attention to who they interact with via email and that this piece of information is important to them as they process new emails.

4.4 Request

The definition of a "phishing" message in this paper is a message (email) that pretends to be something that it is not, in order to get the user to do something they wouldn't normally be willing to do. The second part of that definition is important; phishing isn't just fake email, but it is fake email that requests action.

We wanted to see what kinds of actions were being requested in the suspicious emails that people received and remembered. We asked the respondent whether the email was asking them to do any of a common set of actions. The most common action requested was clicking on a link, which was requested in 57% of the emails reported. This is unsurprising, as this is the stereotypical phishing email, though if anything the surprise was that 40+% of respondents did not remember a requesting link. Only 19% of emails reported asked the user to open an attachment.

46% of emails asked the recipient to respond to the email with some kind of information. That is, rather than using a webpage to collect information or attaching malicious code to the email, the email asked for a response. Responding to emails is a very normal, everyday activity. As an example, consider this story:

P20 Story: *I got an email and it was from an unknowns sender and it was from a different country. As for the country I am unsure of what country it came from. I did not recognize the sender at all. They told me that I won some type of lottery and that all I needed to do was verify my name address date of birth and I could get the money. Then they also said in order to get paid the money all I had to was verify the information and then they would send me the money into my bank account. Then in order for them to send it they needed me to provide them my bank account information my routing number and account number and the banks name and address. I found all of this very concerning and was always told to never give out my social security number or any other personal information to anyone asking for it.*

Almost a third of emails, or 32% of emails, asked the user to take some sort of action outside of the context of email. P39 was asked to make a phone call, for example:

P39 Story: *After receiving a fraud alert email requesting me to call a company I do business with, I checked the phone number, and it was not what I had on file. I also was unaware of any fraudulent activities involving me; however I had my doubts. Therefore I called the number requested, and they started to ask me questions to corroborate my identity. I was reluctant to provide any information, and they told me that*

they would not provide information to me because they were concerned about my identity.

After a bit of a discussion I terminated the call. Subsequently I called the firm at a number that was familiar to me. They wound up transfering me to the fraus department internally. The end result is that the email was legitimate, just poorly constructed). The good news is that therre was no fraus regarding my account.

Most summaries of phishing focus on technical means of information extraction (malicious links, malware attachments) [32], but this suggests that we should also examine non-technical means like simply replying to the email. These incidents that ask for responses or actions outside of email are important reminders that email is a small piece of much larger systems of work, and that email can often be a thing that triggers other types of work to be done. Anti-phishing systems cannot just focus on email; they also need to watch the other non-email work that people do in response to email.

Interestingly, 94% of respondents were able to identify at least one requested action by the suspicious emails. Requesting actions is part of the definition of phishing because it is these actions that the attackers are most interested in. It is good news that users seem to be quite attentive to what actions are being requested, which means this is something that is necessarily present in all phishing emails, and also something that users are good at identifying, which makes it a good place to focus training.

4.5 Suspecting

Our definition of phishing includes that the email is fraudulent — it either explicitly lies or lies by omission about some important aspect of the email. In order to become suspicious of the email, though, it isn't enough to just notice those aspects of the email. The recipient of the email also has to suspect that something is not right about the email.

We asked respondents about each part of the email and whether it felt normal or whether it felt "off" in some way. 59% of respondents reported that the subject line of the email felt "off" in some way. 70% of respondents reported that the sender information felt "off", and 75% of respondents said that the body of the email was "off" in some way. This suggests that all three aspects of an email can provide important clues to end users that an email might be phishing, though the body (content) of an email tends to help users more.

When a respondent felt that the sender was off, they were about twice as likely to indicate that the email address felt off than they were to indicate that the sender's name was the thing that felt wrong. Though, as P99's story shows, the name can also be important:

P99 Story: *Upon strolling through my email account I notice this bogus looking email from what should have been Social Security Administration.*

Except the administration was replaced with bureau & immediately I knew it was bogus.I politely pulled the lil trashcan up for a good old fashion delete session. I usually don't open up anything deemed be to good to be true or bogus or otherwise.

When a respondent felt that the body of the mail felt off, we provided a number of options to them for indicating what in the body felt off. 32% of respondents indicated that the body included unexpected typos or other similar issues. 28% indicated that the body included something strange that isn't normally seen in emails like this. These two aspects suggest that typos are definitely triggers for suspicion, but other strange aspects of emails are almost as common as a trigger.

15% indicated that the email was missing something important. 14% indicated that the email included less information than they would expect. And only 7% indicated that the email included more information than they would expect. To our respondents, phishing emails including less information or missing something triggered suspicions much more often than including too much information. This means that for non-expert end users, their expectations for how much information the emails in their inbox typically include is an important aspect of suspecting an email might be phishing.

4.6 Investigating

Wash [34] points out that people rarely go directly from treating an email as a real email to believing that it is a phishing email. Instead, there is an intermediate stage of "suspicion." When a person is suspicious of the email, they are not sure whether it is legitimate or fraudulent. During this suspicious stage, Wash [34] describes people as taking investigative steps to figure out whether the email is legitimate or not.

We asked respondents about the investigations that they did of their suspicious email. 24% of respondents indicated that they did not do any kind of investigation, and an additional 3% did not remember if they did. That means that 73% of respondents undertook at least one extra step to investigate the email to determine if it was legitimate or not.

The most common investigative step taken was to look more closely at the email address. 36% of respondents in this study indicated that they did this. Looking at the email address seems to be an important everyday step that non-expert users try when they are suspicious of an email.

P66 Story: *An email came in from Paypal describing that a subscription had been purchased with the amount and name of the company/person. I have never seen or heard of the indicated party and at first thought, it may have been a legitimate email. After debating to click the link to login to Paypal and stop the transaction, I hovered over the sender's information and saw the email address had absolutely nothing to do with PayPal's contact information.*

Only 12% of respondents indicated that they looked more closely at a link the email. 7% hovered over the link to see where it went, and 5% actually clicked on the link to see where it went. Link investigation is often mentioned in much phishing training, and it is disappointing that only 12% of respondents investigate links. It is especially disappointing that over a third of those respondents clicked the link as the investigative step.

On the other hand, 16% of respondents reported looking at the headers of the email. This was more common than we expected.

4.6.1 Investigating outside of the email

As mentioned above, emails are frequently just small parts of larger systems. During the investigation, it is possible to look outside of the email for additional information that can inform the decision. In one common method, 18% of respondents reported seeking out a second opinion about the email and asked someone else.

We specifically asked respondents about steps they took to learn more about the purported sender of the email. 82% of respondents reported that they did not take any steps to learn more about the sender, but the remaining 18% did. 9% went to the purported sender's website to get more information about the email. 6% tried to contact the sender via phone. And 1% talked to the sender face-to-face, such as P220:

P220 Story: *I got an email from my work email account from what I thought was my coworker. The body of the email was worded strangely and asked me to click on a suspicious link. I looked closely at the email address it was sent from and it was not exactly correct given my work email addresses. I went to who I thought was the sender face-to-face and asked if he sent the email. He said no and I went ahead and deleted the email.*

Too much phishing training focuses on teaching people to investigate suspicious emails by looking at features internal to the email, such as the sender's email address and links [19,30]. It is surprising that as many as 18% of our respondents took investigative steps outside of the email.

4.7 Deciding

Wash [34] found that after investigating the email, his expert participants would frequently come to a final decision about whether the email was legitimate or phishing. We asked our respondents whether they did come to a final decision, and if so, what that decision was. 80% of respondents did come to a final decision, and almost all of them decided that the email was definitely not safe (78% not safe, 2% safe). The remaining 20% were either still not sure (17%) or don't remember if they came to a decision (3%).

We asked respondents how confident they were in their final decision on a scale of 0 to 10. 69% of respondents chose the highest confidence option (10), and the average confidence was 8.9. Respondents reported very high levels of confidence in their decision about whether the email was safe or not.

4.8 Acting

After deciding whether the email is legitimate or phishing, one decision still remains: what should be done about the email? By far, the most common action was simply deleting the email. 78% of respondents reported that they deleted the email and moved on after deciding it was not safe. 32% indicated that they clicked a button in their interface to report the email as spam or as phishing. Only 4% left it in their inbox.

The survey only asked about actions we knew about ahead-of-time. In the manual coding, we were able to code for more actions. 43% of the respondents mentioned deleting the email, and 15% mentioned clicking a button to mark as spam or phish. Additionally, 9% discussed reporting it to authorities in their story in another way, such as calling an IT help desk.

32% explicitly mentioned a "negative action": that they intentionally chose to not do something (like open the email, or respond). These negative actions are often very strongly worded, and respondents seemed to feel strongly about them, often using language describing bad things to justify not doing things in the future. Consider, for example, how P115 justifies not answering phone calls:

P115 Story: *Computer was shut down because of inappropiate access to a potentially dangerous website. I was telling me that i had to pay a fine of $200 to gain access to my computer. I received a phone call about going to a local store to purchase gift cards. I went so far as going to the store to purchase the gift cards and upon checking out. the clerk at the register informed me that I was being scammed and not to buy these cards. In the meantime I had an open line to this scammer, which I promptly hung up on. Upon arriving home I kept getting phone calls from this person, which I never talked with again.*

An additional 9% of respondents reported taking increased precautions in the future, such as installing a virus scanner or being more careful with emails.

People also have emotional reactions to the email. We asked respondents about their experience of a set of emotions, including "nervous," "fear," "terror," "dread," "worry," and "anxiety". All emotions had very low scores, and no emotion averaged higher than 2.2 out of 5. Despite being unsafe, these emails did not evoke strong emotions from our respondents. Past phishing training, especially that derived from Protection Motivation Theory, has used fear appeals to motivate users [5, 20]. Based on this data, phishing emails generally do not lead to strong emotions, and this could explain why fear appeals do not motivate changes in behavior [6].

5 Discussion

5.1 Humans Identify Phishing Differently

Modern email systems involve multiple layers of protection against phishing attacks. Many email senders include checks for phishing as emails get sent. Most email systems include at least one, and often more than one technical system that filters out emails that are believed to be spam or phishing. Many of these systems also label emails as possibly phishing, as a warning to users (e.g., Google's email system [25]). And end users read emails and make legitimacy determinations on their own.

Reason's Swiss Cheese Model of filtering [28] suggests that when there is a chain of filters like this, the filters work best when each filter works on different principles or using different information than other filters in the chain. If two filters use the same information (e.g. sender from email address) in similar ways, then the holes in the cheese line up and malicious emails that get through one filter are also likely to get through the other. However, if two filters use different information, or operate on the information in fundamentally different ways, then each filter is likely to catch messages that the other filter misses, and including both filters makes the system more resilient to attacks than only including one.

In this paper, we present evidence that this final filter – humans reading emails and determining if an email is legitimate – operates in a very different way, using different knowledge and capabilities, than almost all of the technical filters. We found that humans possess important information that technical phishing filters do not have. They rely on their familiarity with related emails received in the past (72%) and their expectations of incoming emails (95%) to make sense of and become suspicious of phishing emails. This knowledge is highly contextual and very unique to each individual and their experiences. In addition, humans use their knowledge of what was typical in emails they received in the past to spot unexpected and missing important pieces of information in new emails. This information is critical for detecting zero-day phishing attacks, which technical solutions rarely detect [12].

Our respondents were able to notice the nature of the email (e.g. 78% noticed it was personal) and the email account in which the email was received. This requires knowledge of all email accounts a person has and the kinds of communications expected in each account based on how and what the person chooses to use each account for. It is very complex and challenging for technical filters to acquire such knowledge and apply it accordingly, lest they surveil individuals.

Second, we found that humans possess unique capabilities that they use to identify phishing messages, which technical filters do not have. 94% of the non-expert respondents were able of identify what action the email was asking them to do, and over three quarters said they explicitly noticed this about the email. Requests for action are not commonly part of many spam and phishing filters, and when they are, they are often limited in scope mostly by language issues (e.g. checking if the email contains a link to a login page and verifying if the login page is legitimate [23]). Even non-experts are highly attuned to these requests and can confidently identify them.

When filtering, humans also have an investigative ability that technical filters lack: they can choose to take additional time and look up more information from third party sources. A number of our respondents indicated that they would ask colleagues for advice or try to contact the purported sender of the email.

The above are capabilities and knowledge that humans have, but technical phishing filters lack. Following the logic of the Swiss Cheese Model, relying on both humans and technical filtering in combination is better than just relying on one or the other. In recent years, organizations have been relying more heavily on automated phishing detection. Our findings suggest that reducing the diversity of filters may leave systems vulnerable to phishing, and that approaching end user training differently could strengthen strategies for preventing harm from phishing.

Much of the advice about phishing in the IT community involves preventing messages from ever getting to end users [14], rather than trying to educate end users. Because end users are able to filter messages in fundamentally different ways than technical filters, it would be more valuable to spend some money and resources improving the ability of end users to have a significant role in detecting phishing messages. Too much phishing training focuses on technical details (like url parsing [19, 30]) or behavioral changes (like not clicking [20, 35]), instead of trying to strengthen the capabilities that are unique to humans. In this paper, we have presented evidence of some of the knowledge and capabilities that humans have which can be leveraged to enhance phishing training and detection, e.g. forming expectations for emails and asking other people for information.

As the Swiss Cheese Model points out, in a series of filters, putting all of your resources into one layer of filters in exclusion to others removes the benefits you get from a defense in depth strategy. It is often better to have two imperfect filters that operate on different principles or information than it is to have one filter that is highly optimized but limited.

5.2 Similar to Expert Phishing Detection?

Our findings also have implications for identifying similarities between expert and non-expert user phishing email detection. Wash [34] conducted a detailed study of how people detect phishing emails. That study was conducted with IT experts – people with IT training and professional experience that allows them to successfully detect phishing emails. We extended that model, and based many of our questions on that extended model, partially to try to determine if features of that model are also present in how non-experts detect phishing.

In this paper, we are able to validate parts of his model with a non-expert population. Wash also pointed out that in addition to IT expertise, being a knowledge worker can provide expertise in managing email that is relevant to phishing detection. Our sample is not IT experts, and it is also not primarily knowledge workers who deal with email constantly.

In particular, we are able to validate that non-experts do have expectations about what should be present in emails and notice when those things are different. We are also able to validate that even in non-experts, people's attention is focused on what the email is requesting that they do; almost everyone in our study was able to identify what request the email was making. We validated that our non-experts self-reported that they frequently had gut feelings that something was off about the emails, helping them become suspicious. We were able to validate that people would frequently take explicit steps to investigate an email that they found to be suspicious. And we were able to validate that non-experts were able to conclusively decide whether an email was a phishing email or not. This lends support to the implication that expertise about one's own email inbox is an important and yet underutilized aspect of phishing detection training.

We were not able to validate all aspects of Wash's model with non-experts. In particular, Wash's model includes a chronological ordering of stages – first sensemaking, then suspicion, then acting. Our study is a survey and was unable to determine a chronological ordering that things happened in, and as such, we are not sure that things necessarily happen for non-experts in the order that Wash proposes.

5.3 Implications for Phishing Prevention

Email users engage in complex investigations of suspicious emails before they determine if the email is phishing, but current training and technologies do not support these investigations. Our findings suggest that phishing training could support user investigations better by encouraging users to delay taking actions until finalizing their investigation and encouraging email users to leverage peer capabilities (such as asking a friend for help). Additionally, companies that send email can provide helpdesk-style support to help users determine if the company actually sent the email to the user. Email clients could better support investigations by including a "help me troubleshoot this email" button, with contextualized suggestions for investigation.

6 Limitations

This paper is about people, their cognition, and how they successfully detect phishing. It is not about phishing emails. A survey is not a good method for collecting underlying ground truth data on the actual phishing emails or detection failures, because of selection bias and imperfect memory.

Recalling a phishing email prompted recollection of a specific instance, allowing the survey to investigate the processes that people use to detect phishing emails in their inbox. The answers we received were only about this one specific incident, and do not necessarily represent other incidents that the person was involved in; however, across respondents, these answers do represent a variety of the types of phishing incidents that non-experts encounter. Past research has focused almost exclusively on detection failures and fixing those failures; we instead look at what is working well in phishing detection and what should be supported.

Since this is a survey, we can only ask detailed questions about things we know about ahead-of-time. We based our survey questions on Wash's investigation of expert phishing detection [34]. We are not able to determine if the non-experts also use additional methods that were not present in Wash's experts. That is, we seek to learn which of these experts' methods are also used by non-experts, but we cannot learn anything about non-expert methods that are unique to non-experts. Therefore, we do not claim that these methods are a comprehensive description of how non-experts identify phishing; instead, we characterize some methods that they do use.

7 Conclusion

Phishing is a cybersecurity threat that many people experience; almost half of the people eligible for our survey could identify at least one specific phishing email that they received. These people have stories about phishing experiences that they can share with others, and we suspect these stories form an important part of how email users learn about phishing.

We found that many of the techniques that experts use to identify phishing [34], such as noticing minor discrepancies, forming expectations about what the email should look like and noticing differences from those expectations, and becoming suspicious and investigating the email more closely, are also present in how non-experts detect phishing emails.

We also found that much of the information that non-experts use when identifying phishing cannot be replicated by technical phishing detection systems. End users know the purpose (business, personal) of email accounts they receive emails at, and pay attention to that fact. They know whether an email is expected, and are able to compare it against other, similar emails they have received in the past (phishing emails often feel familiar). Additionally, these non-experts have investigative abilities, such as delaying responding to emails and asking the sender for confirmation or more information, that technical phishing filters don't possess. Targeting future phishing training at improving the use of this unique knowledge and expanding the use of these abilities is likely to yield improvement in phishing protection.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1714126. We would like to thank Faye Kollig and Abrielle Mason for assistance with coding the stories and copy editing. All members of the MSU BITLab provided valuable feedback on this study and paper.

References

[1] Jeremy Bryans and Budi Arief. Security implications of structure. In *Structure for Dependability: Computer-Based Systems from an Interdisciplinary Perspective*, pages 217–227. Springer, 2006.

[2] Deanna D Caputo, Shari Lawrence Pfleeger, Jesse D Freeman, and M Eric Johnson. Going spear phishing: Exploring embedded training and awareness. *IEEE Security & Privacy*, 12(1):28–38, 2013.

[3] Debra L. Cook, Vijay K. Gurbani, and Michael Daniluk. Phishwish: a simple and stateless phishing filter. *Security and Communication Networks*, 2(1):29–43, 2009.

[4] Lorrie Faith Cranor. Can phishing be foiled? *Scientific American*, 299(6):104–111, 2008.

[5] Nicola Davinson and Elizabeth Sillence. It won't happen to me: Promoting secure behaviour among internet users. *Computers in Human Behavior*, 26(6):1739–1747, 2010.

[6] Julie S. Downs, Mandy Holbrook, and Lorrie Faith Cranor. Behavioral response to phishing risk. In *Proceedings of the Anti-Phishing Working Groups 2nd Annual ECrime Researchers Summit*, eCrime '07, pages 37–44, New York, NY, USA, 2007. Association for Computing Machinery.

[7] Julie S Downs, Mandy B Holbrook, and Lorrie Faith Cranor. Decision strategies and susceptibility to phishing. In *Proceedings of the second symposium on Usable privacy and security*, pages 79–90, 2006.

[8] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. You've been warned: An empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1065–1074, New York, NY, USA, 2008. Association for Computing Machinery.

[9] Ian Fette, Norman Sadeh, and Anthony Tomasic. Learning to detect phishing emails. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 649–656, New York, NY, USA, 2007. Association for Computing Machinery.

[10] Joshua T Goodman, Paul S Rehfuss, Robert L Rounthwaite, Manav Mishra, Geoffrey J Hulten, Kenneth G Richards, Aaron H Averbuch, Anthony P Penta, and Roderict C Deyo. Phishing detection, prevention, and notification, October 16 2012. US Patent 8,291,065.

[11] The Radicati Group. Email statistics report 2019-2023 executive summary. Technical report, The Radicati Group, 2019.

[12] Ryan Heartfield and George Loukas. A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks. *ACM Computing Surveys (CSUR)*, 48(3):1–39, 2015.

[13] Thorsten Holz, Christian Gorecki, Konrad Rieck, and Felix C Freiling. Measuring and detecting fast-flux service networks. In *The Network and Distributed System Security Symposium (NDSS)*, 2008.

[14] Jason Hong. The state of phishing attacks. *Communications of the ACM*, 55(1):74, Jan 2012.

[15] Scott D Johnson, Jeffrey W Flesher, and Shih-Ping Chung. Understanding troubleshooting styles to improve training methods. In *American Vocational Association Convention*. ERIC, Dec 1995.

[16] Y. Joshi, S. Saklikar, D. Das, and S. Saha. Phishguard: A browser plug-in for protection from phishing. In *2008 2nd International Conference on Internet Multimedia Services Architecture and Applications*, pages 1–6, 2008.

[17] Mahmoud Khonji, Youssef Iraqi, and Andrew Jones. Phishing detection: a literature survey. *IEEE Communications Surveys & Tutorials*, 15(4):2091–2121, 2013.

[18] Gary Klein. *Sources of Power: How People Make Decisions*. MIT Press, 1998.

[19] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Teaching johnny not to fall for phish. *ACM Transactions on Internet Technology (TOIT)*, 10(2):1–31, 2010.

[20] Robert LaRose, Nora J. Rifon, and Richard Enbody. Promoting personal responsibility for internet safety. *Communications of the ACM*, 51(3):71–76, March 2008.

[21] Eric Lipton, David E Sanger, and Scott Shane. The Perfect Weapon: How Russian Cyberpower Invaded the U.S. *The New York Times*, dec 2016.

[22] MacEwan University. University Discovers Online Fraud. Press Release, 2017. https://www.macewan.ca/wcm/MacEwanNews/PHISHING_ATTACK.

[23] L. A. T. Nguyen, B. L. To, H. K. Nguyen, and M. H. Nguyen. A novel approach for phishing detection using url-based heuristic. In *2014 International Conference on Computing, Management and Telecommunications (ComManTel)*, pages 298–303, 2014.

[24] US Bureau of Labor Statistics. Employment–population ratio, Retrieved Feb, 2021. https://www.bls.gov/charts/employment-situation/employment-population-ratio.htm.

[25] Rob Pegoraro. We keep falling for phishing emails, and google just revealed why. *Fast Company*, 2019. https://www.fastcompany.com/90387855/we-keep-falling-for-phishing-emails-and-google-just-revealed-why.

[26] Justin Petelka, Yixin Zou, and Florian Schaub. Put your warning where your link is: Improving and evaluating email phishing warnings. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–15, New York, NY, USA, 2019. Association for Computing Machinery.

[27] Emilee Rader, Rick Wash, and Brandon Brooks. Stories as informal lessons about security. In *Proceedings of the Eighth Symposium on Usable Privacy and Security (SOUPS)*, pages 1–17, 2012.

[28] James Reason. *Human Error*. Cambridge University Press, 1990.

[29] Ozgur Koray Sahingoz, Ebubekir Buber, Onder Demir, and Banu Diri. Machine learning based phishing detection from urls. *Expert Systems with Applications*, 117:345 – 357, 2019.

[30] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the 3rd Symposium on Usable Privacy and Security (SOUPS)*, pages 88–99, 2007.

[31] Rebecca Smith. How a U.S. Utility Got Hacked. *Wall Street Journal*, Dec 2016.

[32] Symantec. Internet Security Threat Report. Technical Report February, 2019.

[33] Verizon. 2019 Data Breach Investigations Report. Technical report, 2019.

[34] Rick Wash. How experts detect phishing scam emails. *Proceedings of the ACM: Human Computer Interaction*, CSCW(160), October 2020.

[35] Rick Wash and Molly M Cooper. Who provides phishing training? facts, stories, and people like me. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.

[36] Steve Whittaker, Victoria Bellotti, and Jacek Gwizdka. Email in personal information management. *Communications of the ACM*, 49(1):68–73, January 2006.

[37] Weining Yang, Aiping Xiong, Jing Chen, Robert W. Proctor, and Ninghui Li. Use of phishing training to improve security warning compliance: Evidence from a field experiment. In *Proceedings of the Hot Topics in Science of Security: Symposium and Bootcamp*, HoTSoS, pages 52–61, New York, NY, USA, 2017. Association for Computing Machinery.

A Survey Instrument

A.1 Consent Form

Thank you for your interest in this research study. After reviewing the consent form below, please select the "I Agree" button if you would like to participate.

What is the purpose of this study? You are being asked to participate in a research study that is being conducted by Dr. Rick Wash and members of the Behavior, Information and Technology Lab (BITLab) at Michigan State University. The purpose of this study is to better understand how people think about and react to email messages they receive that seem suspicious or potentially harmful. You must be 18 years old to participate in this study.

What will I do if I choose to be in this study? Completing this survey should take approximately 20 minutes. The survey consists of multiple choice and fill in the blank questions. You will be asked questions about yourself, and about email messages that you have received. You will then be asked to remember specifics about a suspicious or potentially harmful email message you received in the past, and answer questions about that particular email message.

What are my rights as a participant in this study? You have the right to stop participating at any time. Your decision regarding participating will have no adverse consequences. You have the right to contact the researchers to ask questions about the purposes and procedures of this research after you have finished the survey. You may request that any information you give be ignored, or that any or all data from your survey be destroyed.

What are the risks and benefits of participating? Your participation in this study does not involve any physical or emotional risk to you beyond that of normal, everyday use of the Internet and email. You may not directly benefit from your participation in this study. However, your participation in this study may contribute to the understanding of how people think about suspicious email messages they receive. This will help researchers to develop tools and training that could prevent email messages from causing harm in the future.

How will I be compensated? If you successfully complete the entire survey, you will receive the incentive stated in your invitation in return for your participation.

What about the confidentiality and privacy of my information? Your survey responses will be assigned an anonymous code number, and researchers will save all survey responses by this code number. Any personally identifying information that you may provide in your answers to the survey questions will be removed by researchers before analyzing the data, so your answers cannot be linked with your name or identity in any way.

Survey responses and aggregate results of this research may be used for teaching, research, publications, or presentations at professional or scientific meetings. They may also be used for future research studies or shared with other researchers for secondary analysis or use in other research without additional informed consent from you. This means researchers may publish, present and share with other researchers summaries of data from multiple people, and direct quotations from individual responses.

No potentially sensitive, incriminating, or identifying information about you or others mentioned in the survey responses will be used in any publication or presentation, or shared outside the research team, except as required by Michigan State University's Human Research Protection Program or by law. Any use of your responses for public consumption will be carefully anonymized so it does not contain any identifying information.

Please note that the data will be retained at Michigan State University for a minimum of 5 years after all analyses and publications related to this project have been completed. Data will be stored on a secure, password-protected computer.

Whom should I contact if I have questions or concerns about this research study? If you have concerns or questions about this study you may contact Dr. Rick Wash, who is in charge of this research study, at telephone number 517-355-2381 or by email at wash@msu.edu.

If you have questions or concerns about your role and rights as a research participant, would like to obtain information or offer input, or would like to register a complaint about this study, you may contact, anonymously if you wish, the Michigan State University's Human Research Protection Program at 517-355-2180, Fax 517-432-4503, or e-mail irb@msu.edu or regular mail at 4000 Collins Rd, Suite 136, Lansing, MI 48910.

Consent to participate Your participation in this study is completely voluntary. By clicking "I agree" below you are voluntarily agreeing to participate.

Q: Please select "I agree" below if you would like to participate.
○ I agree
○ I do not agree

A.2 Screening

Q: Have you ever received formal training in computer science, software engineering, IT, computer networks, or a related technical field?
○ Yes
○ No
○ I'm not sure

Q: Have you ever worked in a "high tech" job such as computer programming, IT, or computer networking?
○ Yes
○ No
○ I'm not sure

Q: What is your age in years?

Q: In this survey, we are interested in hearing about emails you received that were suspicious or potentially harmful in some way. This can be any email that you were suspicious about, including emails that you were concerned about but ended up not being a problem.

We are very interested in hearing about emails where it was hard for you to figure out what to do. For example, this could be an email message that you were unsure of and had to look closely at it to figure out if it could be harmful. Many of these emails ask you to do something, like click a link, open an attachment, or respond to the email with information.

Can you remember any suspicious or potentially harmful email messages that you've received? It is OK to go look through your email account and then continue with the survey, to help you recall if you've ever received email messages like this.
○ Yes, I have received email messages like this in the past.
○ No, I do not remember receiving any email messages like this.
○ I'm not sure

Q: What gender do you identify as?
○ Man
○ Woman
○ Other (fill in the blank)
○ Prefer not to answer

Q: Which categories below best describe you? Select all that apply:
☐ White
☐ Hispanic, Latino or Spanish
☐ Black or African American
☐ Asian
☐ American Indian or Alaska Native
☐ Middle Eastern or North African
☐ Native Hawaiian or Other Pacific Islander
☐ Some Other Race, Ethnicity or Origin (please specify)

A.3 Elicitation

Q: First, to help you to remember emails that were suspicious or potentially harmful, please list some different ways that an email message can be unsafe or cause security problems:

Q: Next, think about different ways you know of to recognize an email that is suspicious or potentially harmful, and make a list of these below:

Q: Take a moment to think about times in the past when you personally received a suspicious or potentially harmful email. Please list as many of these emails as you can remember, using only a couple of words to describe each one. You may want to re-read your answers to the previous questions to jog your memory.

Q: On the previous page, you made a list of emails that you personally received that were suspicious or potentially harmful. For reference, here is the list:

Q: Choose one email message from the list above that it's easy for you to recall details about. You will be answering questions about this email in the rest of the survey. Briefly summarize that email, and what happened when you received it.

Q: In approximately 5 words, please describe what made this email seem suspicious:

Q: In approximately 5 words, please describe why it was hard for you to figure out how to deal with this email:

Q: In approximately 5 words, please describe what the email was asking you to do:

A.4 Noticing

For your reference, here is what you said about the email you will be answering questions about on this page:

Q: How long ago did you receive the email?
○ Within the last day
○ Within the last week
○ Within the last month
○ Within the last year
○ Longer than one year ago
○ I don't remember

Q: To help us monitor the quality of our data, please select "Somewhat disagree" from the choices below.
○ Strongly agree
○ Somewhat agree
○ Neither agree nor disagree
○ Somewhat disagree
○ Strongly disagree

Q: At which of your email accounts did you receive the email?
○ Work Email account
○ Student Email account
○ Personal Email account
○ Other (please describe)
○ I don't remember

Q: What was the context of the email?
○ This email was related to work
○ This email was of a personal nature
○ Other (please describe, briefly)
○ I don't remember

Q: Who did the email appear to come from?
○ A work colleague
○ A close friend or family member
○ An acquaintance from outside work
○ A company, business or other organization

○ Other (please describe)
○ I don't remember

A.5 Expecting

For your reference, here is what you said about the email you will be answering questions about on this page:

Q: Please indicate your agreement or disagreement with the following statement:

When I read the email message, I felt like I had received other email messages like this one before.
○ Strongly agree
○ Somewhat agree
○ Neither agree nor disagree
○ Somewhat disagree
○ Strongly disagree

Q: Before receiving this email, had you ever received an email message from the sender?
○ Yes
○ I'm not sure
○ No
○ I don't remember

Q: Before receiving this email, had you ever interacted with the sender in some other way than email? (For example, if you had previously talked to the sender face-to-face, or visited their website.)
○ Yes
○ I'm not sure
○ No
○ I don't remember

Q: Before receiving this email, how long had you known the sender?
○ One month or less
○ Between one month and one year
○ One to two years
○ Two to five years
○ Five to ten years
○ More than 10 years
○ I don't remember
○ I did not know the sender

Q: Did you expect to receive this specific email?
○ Yes
○ I'm not sure
○ No
○ I don't remember

Q: Please indicate your agreement or disagreement with the following statement:

This email message seemed different from the email messages I typically receive.
○ Strongly agree
○ Somewhat agree
○ Neither agree nor disagree
○ Somewhat disagree
○ Strongly disagree

A.6 Suspecting

For your reference, here is what you said about the email you will be answering questions about on this page:

Q: Many suspicious emails ask you to do something. Was the email asking you to do any of the following? Please check all that apply.
☐ Click on a link or button
☐ Open something that was attached to the email
☐ Respond to the email with some information
☐ Take some action outside of the email
☐ None of the above
☐ I don't remember

Q: Think about the subject line of the email. Did the subject line feel normal, or did it feel "off" in some way?
○ I didn't notice anything that felt off about the subject line
○ The subject line was different than I would expect
○ I don't remember much about the subject line of the email

Q: Think about who the email said it was from. Did this sender information make sense, or did it feel "off" in some way?
○ I didn't notice anything that felt off about the sender
○ The sender's name looked different than I would expect
○ The sender's email address looked different than I would expect
○ I don't remember who the email said it was from

Q: Think about the main body of the email. Did the main body of the email seem normal, or did you notice anything that felt "off" about it? Please check all that apply.
☐ I didn't notice anything that felt off about the main body of the email
☐ The main body of the email included typos or other issues that I didn't expect to be in an email like this
☐ The main body of the email was missing something that I would expect to be in an email like this
☐ The main body of the email included something strange that I do not normally see in an email like this
☐ The main body of the email included more information than I expect to be in an email like this
☐ The main body of the email included less information than I expect to be in an email like this
☐ I don't remember much about the main body of the email

Q: When you read the email, did you believe that the email was harmful?
○ Yes, I thought it was harmful
○ I was not sure about whether it was harmful or not
○ No, I did not think it was harmful

○ I don't remember

Q: How sure or unsure are you about your answer to the previous question?

Please indicate your answer below on a scale from 0-100, where 0 means COMPLETELY UNSURE and 100 means COMPLETELY SURE.

A.7 Investigating, Deciding, and Acting

For your reference, here is what you said about the email you will be answering questions about on this page:

Q: What actions did you take to learn more about the email? Please check all that apply.
☐ Hovered over one or more of the links in the email to see where it went
☐ Clicked on one or more of the links to see where it went
☐ Looked more closely at the the email address the email came from
☐ Opened the attachment
☐ Looked at email headers
☐ Asked someone else about the email
☐ None of the above
☐ I don't remember
☐ Other

Q: In what ways did you attempt to learn about the sender of the email? Please check all that apply.
☐ I went to the website of the sender
☐ I contacted the sender via phone
☐ I contacted the sender through another communications medium (texting, chat, social media)
☐ I talked to the sender face-to-face about the email
☐ I did not try to contact the sender
☐ I don't remember

Q: After you learned more about the email, did you decide that the email was safe or not?
○ Yes, the email was safe
○ I was still not sure whether the email was safe or not
○ No, the email was definitely not safe
○ I don't remember

Q: How sure or unsure are you about your answer to the previous question?

Please indicate your answer below on a scale from 0-100, where 0 means COMPLETELY UNSURE and 100 means COMPLETELY SURE.

Q: What action(s) did you take with this email? Please check all that apply.
☐ Deleted the email
☐ Clicked a button to report the email as spam
☐ Sent the email to someone
☐ Responded to the email
☐ Left the email in my inbox
☐ None of the above
☐ I don't remember

Q: At any point while handling this email, to what extent did you experience these emotions?
Scale:
○ Not at all
○ Somewhat
○ Moderately
○ Quite a bit
○ An extreme amount
Emotions to be rated on that scale:
○ Dread
○ Terror
○ Anxiety
○ Nervous
○ Scared
○ Panic
○ Fear
○ Worry

Q: Please indaicate your agreement or disagreement with the following statement:

I feel like something harmful happened because of this email message.
○ Strongly disagree
○ Somewhat disagree
○ Neither agree nor disagree
○ Somewhat agree
○ Strongly agree

A.8 Full Story

You are almost done! You have now answered a number of questions about an email message that seemed suspicious or potentially harmful, and hopefully you have recalled quite a few important details. For reference, here is the short description of the email you provided at the beginning of the survey:

Q: Below, please write the story of the email as if you were telling it to a friend. Use as much detail as you can, including any thoughts or recollections about what happened you might have had as you were filling out the survey. Your story should be at least 4 or 5 sentences long (minimum 300 characters).

Q: How easy or difficult was it for you to remember a suspicious or potentially harmful email to answer questions about in this survey?
○ Extremely easy
○ Somewhat easy
○ Neither easy nor difficult
○ Somewhat difficult
○ Extremely difficult

A.9 Demographics

Q: What is the last grade or class you completed in school?
○ None, or grades 1-8
○ Some high school
○ High school graduate or GED certificate
○ Technical, trade, or vocational school AFTER high school
○ Some college, no 4-year degree
○ 4-year college degree
○ Some postgraduate or professional schooling, no postgraduate degree
○ Postgraduate or professional degree, including master's, doctorate, medical or law degree

Q: What is your current employment status?
○ Employed full time
○ Employed part time
○ Unemployed looking for work
○ Unemployed not looking for work
○ Retired
○ Student
○ Student and employed part time
○ Disabled

Q: What was your total household income before taxes during the past 12 months?
○ Less than $25,000
○ $25,000 to $34,999
○ $35,000 to $49,999
○ $50,000 to $74,999
○ $75,000 to $99,999
○ $100,000 to $149,999
○ $150,000 to $199,999
○ $200,000 or more

Q: How familiar are you with the following Internet-related terms?

Please rate your understanding of each term below from None (no understanding) to Full (full understanding):

Scale:
○ None
○ Little
○ Some
○ Good
○ Full

Terms to be rated:
○ Wiki
○ Meme
○ Phishing
○ Bookmark
○ Cache
○ SSL
○ AJAX
○ RSS
○ Filtibly

A.10 Thank You

Thank you for participating! If you have any questions or concerns, please contact Dr. Rick Wash at telephone number 517-355-2381 or by email at wash@msu.edu.

Code Reviewing as Methodology for Online Security Studies with Developers – A Case Study with Freelancers on Password Storage

Anastasia Danilova
University of Bonn
danilova@cs.uni-bonn.de

Alena Naiakshina
University of Bonn
naiakshi@cs.uni-bonn.de

Anna Rasgauski
University of Bonn
rasgausk@cs.uni-bonn.de

Matthew Smith
University of Bonn, FKIE Fraunhofer
smith@cs.uni-bonn.de

Abstract

While ample experience with end-user studies exists, only little is known about studies with software developers in a security context. In past research investigating the security behavior of software developers, participants often had to complete programming tasks. However, programming tasks require a large amount of participants' time and effort, which often results in high costs and small sample sizes. We therefore tested a new methodology for security developer studies. In an online study, we asked freelance developers to write code reviews for password-storage code snippets. Since developers often tend to focus on functionality first and security later, similar to end users, we prompted half the participants for security. Although the freelancers indicated that they feel responsible for security, our results showed that they did not focus on security in their code reviews, even in a security-critical task such as password-storage. Almost half the participants wanted to release the insecure code snippets. However, we found that security prompting had a significant effect on the security awareness. To provide further insight into this line of work, we compared our results with similar password-storage studies containing programming tasks, and discussed code reviewing as a new methodology for future security research with developers.

1 Introduction

Code reviewing is a technique applied at the end of the Software Development Life Cycle (SDLC), used as one of the final steps by software developers to ensure programming code quality before software release. Thus, software developers change their perspective from a code creator to a code inspector, which might affect their security awareness. While knowledge on code reviewing in the field of software engineering exists [9, 19, 32], only little is known about this methodology within a security context [13]. Acar et al. [4] called for more studies on developers' security behavior and on the study methodology of security developers. While most of the previous work within this context includes surveys (e.g., [7, 11, 21, 28, 37]), interviews (e.g., [6, 11, 18, 21, 24, 36–38]) or programming tasks (e.g., [2, 3, 17, 20, 22, 23, 25, 25–27, 30, 39–42]), we explored code reviewing as a promising methodology for developer security studies.

Conducting security studies with software developers can be difficult due to recruitment and study compensation challenges [3–5, 10, 12, 20–22, 25, 31, 42]. Programming tasks can also often take more time than professionals can afford. Naiakshina et al. [22–25] conducted a number of studies where computer science (CS) students, freelance developers and software developers from companies were required to complete the registration functionality in a web application. The authors investigated participants' security behavior with a focus on the storage of user passwords in a database. According to the study design and participants' feedback, the study lasted around 8 hours. Recruiting a high number of employed developers who had time outside of their normal working hours for a one-day study was reported to be extremely difficult. Thus, their sample size is not as large as they would have wished.

Therefore, researchers often tend to design programming tasks in such a way that software developers only need to solve small and short tasks (e.g., [2, 3, 5, 33]). For example, Acar et al. [5] conducted a security developer study with GitHub users and provided them with programming tasks which were "short enough so that the uncompensated participants would be likely to complete them before losing interest, but still complex enough to be interesting and allow for some mistakes." While Acar et al. did not mention security or privacy in the recruitment message at all, Naiakshina et al. tested if

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Vancouver, B.C., Canada.

explicitly asking participants for secure password storage (security prompting) would affect their solutions. They found that security prompting had a significant effect on participants solutions. Additionally, participants tended to concentrate on the functionality of the software first, before working on security aspects [22–24].

To provide deeper insights into this research field, we tested code reviewing as a promising methodology for security studies with developers. Instead of asking developers to program a piece of code, we showed them functional code snippets and asked them to write code reviews about the snippets. We based our code snippets on the participants' submissions from the previous freelancer study from Naiakshina et al. [23] and also recruited freelance developers. This allowed us to compare the results of the code review task with the findings of the previous study containing a programming task. While we offer insights on freelancers' behavior in code reviewing tasks on a primary level, we also discuss code reviewing as a methodology for developer security studies on a meta-level. Our main research questions are as follows:

RQ1: How do developers behave when reviewing code in a security-critical task such as password storage? We were interested to find which criteria the developers mostly base their reviews on in security-critical code, and whether they would be able to indicate the security issue, even when being presented with distraction tasks not related to security. Additionally, how detailed would their security problem description and suggestions for improvement be? Would they suggest to release the code even though it contains security-critical issues?

RQ2: Which factors have an influence on developers' security awareness? In particular, we investigated whether prompting, programming experience or code snippets with different password-storage issues (plain text, Base64, MD5) have an influence on whether developers find the security issues.

RQ3: How much time do developers dedicate to security and do they feel responsible for security in a code review? We asked the freelance developers to indicate how much time they spend with security in code reviews and whether they feel responsible for security. These results were compared to their self-reported responsibility and time spent on security in programming tasks.

RQ4: Comparing the results of a programming and a code reviewing task on password storage, which methodological implications can we conclude? To provide insights for security studies with developers, we discuss the advantages and disadvantages of code reviewing as a study methodology compared to programming tasks.

2 Related Work

Our related work section is divided into two parts. First, we discuss related work in the area of developer, security and password-storage studies. We then take a closer look at code-reviewing studies.

2.1 Developer and Password Studies

A qualitative study from Naiakshina et al. [24] in 2017 investigated the password-storage implementation behavior of 20 CS students. The participants were given a scenario where they were told they were working in a team to implement the registration functionality of a social networking platform. Half of the students were prompted beforehand and told to implement a secure solution. None of the non-prompted students implemented a secure solution. As often found with end users, there was also a tendency with developers to offer functionality over security. This study was extended by the authors in 2018, by inviting 20 additional CS students [25]. The authors acknowledged that they had challenges to recruit enough participants for the study. Out of a pool of 1600 CS students at their university, only 40 participated in the study. The exploratory quantitative study supported the findings of the qualitative study. However, CS students stated that they would have implemented a secure solution had the task been for a real company.

To find out whether the previous results were a study artifact, in 2019 Naiakshina et al. [23] conducted a similar study, but this time with 43 freelancers. Hired through Freelancer.com, the participants were asked to implement the registration functionality for a fictitious online sports-picture sharing platform. Here it was also found that prompting has an effect on the security of the end solutions. 15 of the 22 non-prompted, and even 3 of the 21 prompted freelancers received security requests after submitting insecure solutions [23]. Unlike the previous study, the participants believed they were working for a real company. This did not appear to have an impact on the security of the end solution. The same study was also conducted with 36 professional developers from diverse companies in 2020 [22] and again, prompting had an effect on the security of participants' submissions.

Another study on password-storage was conducted by Wijayarathna et al. [40]. Their study was similar to the one described in the previous papers. To explore the usability of the SCrypt password hashing functionality of Bouncycastle, the authors conducted a 2-hour study with 10 programmers and reported 63 usability issues with the Bouncycastle API. A further study from Wijayarathna et al. [39] investigated how security responsible programmers feel when writing code. The participants were expected to complete four programming tasks including a password storage task. They found that developers know they are responsible for security, but often have a difficult time implementing security measures.

A study investigating having GitHub members for developer studies was conducted by Acar et al. [5]. 307 participants completed three small python programming tasks and then filled out a survey, all without payment. One of the tasks was to store login credentials in a database. The researchers set out to explore how well GitHub users can replace IT-professionals in developer studies, who often do not have time for studies outside of their working schedules and who often have an hourly rate much higher than that offered for taking part in studies. They found that whether a participant is a professional or a student had no significant effect on the security perception of the participant or on the functionality or security of the solution implemented. Experience did however have an effect: each year of added python experience increased the chance of a secure solution by 5%.

Tahei et al. [34] reviewed literature looking at security from a developer perspective. They found there was generally a lack of research in developer-centered-security, and that security needed to become of higher business value. As found in the previously mentioned studies, security knowledge among software developers is often lacking and security is often a secondary requirement.

2.2 Code-reviewing Studies

Baum et al. [9] looked at code-reviewing practices in 19 firms, 11 of which regularly conduct some kind of code review. They found that some firms use multiple reviewers, particularly when making large changes to code. Aside from being used to improve the quality of the code and find defects, the researchers found that code reviews are also conducted to enable a learning process of the coder. Furthermore, it was found that some firms do not conduct code reviews: developers can often feel attacked by code reviews and need time to adjust to the new method. Bacchelli et al. [8] supported the finding that code reviews are often used as a learning process in firms. They also found that "finding defects" in code was a driving factor for conducting code reviews. In multiple papers the term "finding defects" included security issues, but mostly included other kinds of defects [8, 9, 29].

Using the largely open source Mozilla project, Kononenko et al. [19] studied the quality of code-reviews. 54% of bugs were not found during the code-review. The authors found a positive correlation between reviewing experience and number of bugs found. Coding experience did not have an impact on the number of bugs found. They found larger patches and an increased number of files to be reviewed increased the chance of introducing "buggy" patches. The use of super reviewers, a reviewer who is a highly experienced developer, decreases the probability of introducing patches with bugs in them.

What effect availability bias has on code reviews was investigated by Spadini et al. [32]. Using a browser-based reviewing tool chosen by the researchers, the participants were expected to review a code change. The change contained three errors: two errors were of the same type and were bugs that reviewers do not usually look for, whereas as the third was of another type. Half of the participants were non-primed and received a code change where none of the bugs had been commented on, whereas as the other half were primed and received a code change where one of the errors had been commented on by another reviewer. The researchers found a correlation between priming the participants and the participants finding the other error of this type. 80% of participants mentioned they were influenced by the comments in the code. The non-primed bug was found by both groups at a similar rate.

A study investigating the optimal number of code-reviewers for a given task was conducted by Edmunson at al. [13]. The participants were web developers all with various backgrounds and security experience. They were presented with and expected to review an existing open-source project, which had some security issues. Edmunson et al. found a correlation between the number of correct vulnerabilities found and the number of false vulnerabilities found. Having more than 15 reviewers did not bring any further improvement. Interestingly, the researchers found a negative correlation between the number of years of experience and the number of correct vulnerabilities found. Our study differs from Edmunson et al.'s study in several aspects. First, while Edmunson at al.'s participants were informed about security issues within the web application they had to review, we did not mention that our code had issues at all. Additionally, half of our participants were advised to consider security for the code review, while the other half were not. Second, in addition to the security issues, we also added general issues within the code, so we were able to see which issues participants are more aware of, if they recognized them at all. Third, we showed participants one simple code snippet in the context of secure password storage, whereas Edmunson et al. investigated Cross-Site Scripting and SQL injection vulnerabilities within an entire web application project.

Most work using code reviewing as a study methodology was conducted in the field of software engineering. It is unclear yet, whether the findings are transferable to a security context. For example, Kononenko et al. [19] found a positive correlation between reviewing experience and number of bugs found, unlike the study from Edmunson et al. [13]. With our study we aim to provide deeper insights into code reviewing as a methodology for developer studies within a security context.

3 Methodology

Past work showed that software developers perceive security as more of a secondary task during programming and thus need to be explicitly asked to consider security aspects in the task requirements [22–24, 27]. It is however unclear, whether

this holds true for code reviewing. The fact that code reviews are usually written at the end of the SDLC might affect developers' security awareness and thus improve the quality and security of the code.

In this work, we investigated whether software developers think of security when writing a review for security critical code, such as user password storage. For this we set up an online survey, for which we recruited freelance developers on Fiverr.com [1]. The complete survey can be found in the Appendix A. Half the participants were prompted for password-storage security prior to writing the review, and the other half were asked without being prompted to write the review. Hence, we explored the independent variable (IV) *security prompting* with the two values prompting and non-prompting. Additionally, we investigated whether different password storage implementations affect the code reviews. Thus, the participants were shown at random one of three insecure code snippets (plain text, Base64, MD5). Hence, our second IV variable was the *code snippet* each participant received, leaving us with a total of six conditions within our study. We conducted a between-subjects study, where each participant randomly received one of the three snippets.

Apart from prompting the prompted groups of participants to ensure the password is stored securely, we did not give any criteria to complete the review. We wanted to find out which criteria the participants chose and which issues they found without further requests. We recorded all questions received from the participants during the task and recorded our answers to these in a play-book to avoid giving more information to some participants than to others (see Appendix B). In addition to that, we provide further insights on a meta-level for using code reviewing as a new, promising methodology for developer studies within a security context.

We conducted a pilot study with one participant to test the survey and to get a better time estimation for the task. The participant finished within two hours. After correcting minor issues with the survey, we conducted the actual study between May and July 2020. We asked our participants to complete the code reviewing task within one week. Thus, we hoped to increase the number of participants.

3.1 Survey

In the survey, we showed the participants one of three code snippets, each of which contained a password storage implementation. The participants were asked to write code reviews for the snippets. We also asked them to list criteria on which they based their reviews and if they would release the code as it is, with minor adjustments or not at all. After the participants finished their code review, they were requested to explain the concepts of hashing and salting. Further, we asked them about their code reviewing experience and how much of a priority security is to them.

We switched off the back button to prevent the participants from changing their code reviews after being asked for code security. That way we aimed to get an unbiased view on the code reviews and to avoid priming participants for security by the survey.

Since we did not ask participants to write but only review programming code, we included a small programming test to the survey, which was also used by Danilova et al. in [11]. Danilova et al. recruited software developers for an online survey study on security warnings. To assure that their participants really had programming skills, the authors designed a multiple choice question with a code snippet where "hello world" was printed out backwards. About 74% of the participants recruited online on the survey platform Qualtrics failed the test, although all of them indicated to have programming skills. To ensure data quality, our participants were also shown this multiple choice question. Finally, the participants had to answer demographic questions.

3.2 Code Snippets

For the code review task, we chose code snippets submitted by freelancers from the study conducted by Naiakshina et al. [23] for two reasons. Firstly, it was programming code created by freelance developers who believed they were working on the registration functionality for a real company which was to be submitted for release. Secondly, it allowed us to compare our code review results with the findings of the programming code analysis from Naiakshina et al. with regards to both the effectiveness of security prompting and the accuracy of the submissions.

Additionally, we wanted to test whether different programming code snippets have an influence on our participants' submissions. We concentrated on the bad practices used by freelancers in [23]. While we decided upon a plain text code snippet as a baseline, we also added one snippet using MD5 as a hashing function and another using Base64 encoding, as these were prevalent within the freelancers' submissions in [23]. We made sure that the three snippets had an approximately similar length (120-130 lines). To further improve their comparability, we adjusted the selected snippets in the following way. First, to reduce the risk of comments influencing the code review, we deleted all comments from the chosen snippets. To look more realistic, but still stay comparable, we added generic comments which were the same for all snippets. We added comments to the head of the class as well as in the main functions, but not to the trivial setters and getters. Second, we rearranged the order of the functions so that all participants would see the functions in the same order when reviewing the code, as we did not want the function order to influence our results. Third, we added two distraction tasks to all of the snippets:
1) *Exception swallowing*: An empty catch block is considered to be bad practice as possible exceptions would be ignored.
2) *Logical mistake*: Within an if-loop we used only one "="

Table 1: Demographics of participants (n = 44)

Age	min: 19, max: 35	sd: 3.83	median: 24.0	mean: 25.06
General Programming Experience	min: 1, max: 12	sd: 2.43	median: 4.0	mean: 4.46
Java Experience	min: 0.5, max: 10	sd: 2.3	median: 2	mean: 3.19
Gender	Female: 3	Male: 40	Prefer to self-describe: 1	
Occupation	Freelance Developer: 27 Academic Researcher: 1	Industry Developer: 8 Industry Tester: 1	Undergraduate: 2 Graduate: 1	Other: 3 Freelance Tester: 1
Country of Residence	Pakistan: 21, India: 8 Nigeria: 1, Turkey: 1	UK: 3, Portugal: 1 Malaysia: 1, Italy: 1, US: 1	Burkina Faso: 1 Bangladesh: 1	Morocco: 1 Sri Lanka: 2 , NA: 1

in the condition. The condition is therefore always true since an assignment is executed instead. We expected that since the participants are eager to find mistakes in a study on code reviewing, these distraction tasks could divert the attention of the participants from the password storage implementation. The three code snippets can be found in the Appendix C.

Table 2: Number of participants per group (n = 44)

	Plain text	**MD5**	**Base64**
Prompted	8	7	8
Non-prompted	6	7	8

3.3 Participants

Like Naiakshina et al. [23], we wanted to recruit freelance developers on Freelancer.com. However, our project was repeatedly denied by the platform with generic explanations such as: "Your project shows behavior that is contrary to our Code of Conduct." We contacted the platform's support service several times to clarify that we wanted to conduct a scientific study with freelancers. However, we were not able to solve our issues at that time and thus decided to use another freelancer platform for the recruitment of participants: Fiverr.com [1].[1]

In the freelancer study by Naiakshina et al., participants received either €120 or €220 for participating in a study of six to eight-hours. No significant difference was found on the security of the submissions between the different payment groups. Since our study was estimated to take one hour and the participant in the pilot study needed two, we decided on a compensation of $50.

We posted our project in four iterations receiving up to 15 applications per posting. With each iteration, the number of repeated offers increased. On Fiverr.com it is required to attach categories and subcategories to the post. Our post was included in the categories "Programming and Tech", "Web-programming" and "Java." In total we received 61 applications to take part in our study. All except four were invited to the study; reasons for not being invited to take part included being under 18 or not having any programming experience. Four freelancers did not respond upon our invitation, six did not want to participate, and two participants canceled after invitation before starting with the study. Finally, 45 freelancers completed our survey. To assure data quality, we excluded one participant from our data set who was not able to answer the "hello world" backward question from Danilova et al. [11].

Table 1 summarizes the demographics of our participants. Out of 44 participants, 40 reported to be male, 3 female, 1 preferred to self-describe. They reported to be between 19 and 35 years old (mean: 25.06 years, median (md): 24 years, sd: 3.83). Further, most of the participants reported to live in Pakistan or India. The general programming experience ranged between one and 12 years with a median of 4 years. All except 2 had at least 2 years of general programming experience and all but 8 reported to have at least 2 years of Java experience (min: 0.5, max: 10, md: 2, mean: 3.19). Most (27) named freelancing as their main profession. All except two participants reported to have reviewed code by others in the past. Table 2 shows the number of valid participants in each group.

3.4 Evaluation

3.4.1 Security

We evaluated the security of participants' code review submissions in the following way. First, we introduced a binary variable *found password storage issue* with two values: 1: participants stated in their reviews, that they found some issues with password storage security; 0: participants did not state in their reviews, that they found some issues with password storage security.

Second, to identify how accurate our participants' code reviews were with regard to the secure password-storage parameters, we used the *security score* of Naiakshina et al. in [23]. Participants received 2 points for hashing and salting the user

[1] Another reason the support service of Freelancer.com offered us was: "Academic cheating is not allowed." As pointed out by Naiakshina et al. [23], it seems that students often use this platform for hiring freelancers to do their university homework. We are still in contact with the enterprise department of Freelancer.com, which reassured us that university studies are welcome to use their platform. It seems that an enterprise self-service would have solved our previous issues.

passwords, another 2 points if the salt was randomly generated and at least 32 bits in length, and another 3 points for iterations, a memory-hard hashing function and if the hash's derived length was at least 160 bits long [24]:

1. The end-user password is salted (+1) and hashed (+1).
2. The derived length of the hash is at least 160 bits long (+1).
3. The iteration count for key stretching is at least 1 000 (+0.5) or 10 000 (+1) for `PBKDF2` and at least $2^{10} = 1\,024$ for `bcrypt` (+1).
4. A memory-hard hashing function is used (+1).
5. The salt value is generated randomly (+1).
6. The salt is at least 32 bits in length (+1).

3.4.2 Qualitative

The code reviews were evaluated qualitatively with inductive content analysis [14]. We decided upon an inductive coding method, as opposed to a deductive coding method, as we did not want to assume what our results would be. The evaluation process included open coding and creating categories. Since we used the "independent parallel coding" approach of David R. Thomas [35], we compared two sets of categories and report the inter-coder agreement for them. The two sets of categories were subsequently merged into a combined set. We calculated the inter-coder agreement Cohen's Kappa (κ) and received an agreement of 0.82. Fleiss et al. considered a value above 0.75 a good level of coding agreement [16].

3.4.3 Quantitative

We additionally conducted an exploratory quantitative analysis. We established the variable from the reviews (found password storage issue) and evaluated the effect of our two IVs (security prompting, code snippet) on this variable using Fisher's exact tests (FET) [15, p. 816]. To test for correlations, we used the Pearson's correlation coefficient. To examine effects in continuous data, we used Wilcoxon Rank sum tests. All tests referring to the same dependent variable (found password storage issue) were corrected using the Bonferroni-Holm correction. The corrected p-values are referred to as $cor-p$.

3.5 Ethics

The institutional review board of our university reviewed and approved our project. We provided the participants of our study with a consent form outlining the scope of the study, the data use and retention policies; we also complied with the General Data Protection Regulation (GDPR). The participants were informed of the practices used to process and store their data and that they could withdraw their data during or after the study without any consequences. Also, the participants were asked to download the consent form for their own use and information.

3.6 Limitations

The participants who took part in this study were recruited on Fiverr.com. They may not be representative for all developers and results may even differ among different freelance platforms. Code reviews are usually performed in companies, so the freelancers might not have had experience with code reviews. However, we aimed to test code reviewing as a study methodology for developer studies as opposed to studying the code review experience.

The majority of participants were non-native English speakers and their responses were not always so clear to understand. Some participants may have had trouble understanding the questions. While this is not desirable for a study, it still represents a realistic scenario, since freelancers with the same issues are hired for real life projects. Moreover, all participants were informed that this project is part of a study. It could be that the participants would have behaved differently had they been writing a code review for a real life project.

We conducted an a priori power analysis with an effect size from Naiakshina et al. from [22] to calculate the necessary sample size to prevent type II errors of falsely rejecting null hypotheses. The required sample sized turned out to be 45 persons per group. However, we were not able to recruit enough freelancers on Fiverr.com, even though we used multiple rounds of recruitment and posted the project repeatedly on the platform. The recruitment of software developers is a challenging task and small sample sizes can limit the method's potential and the generalizability of results. Therefore, our analysis needs to be considered as an explanatory first glance on the problem.

4 Results

The results section is structured as following. First, we present the findings of our qualitative analysis. Second, we report the results of our exploratory quantitative analysis. Third, we compare our results with a similar study containing a programming task on password-storage. We report statements of specific participants by labeling them according to their conditions. The first letter of the label refers to **P**rompting or **N**on-Prompting. The second denotes the code snippet used (**B**ase64, **M**D5, or **P**lain text). While qualitative analysis is more frequently used to explore phenomena, we still provide the numbers of participants to give an indication of the frequency and distribution of themes. An overview of the evaluation of participants' submissions can be found in the Appendix D.

4.1 Qualitative Analysis

The reviews of our participants differed in quality, word count and content. The majority of participants (36 of 44) reported to have reviewed the snippet manually, 4 said that they used an IDE (Eclipse, NetBeans, Visual Studio Code) to check the code, and 4 reported to have used a static analyzer (PMD, sonarlint, findbugs, codacy). With such a small sample size, we could not draw conclusions but it might be worth exploring the use of static analyzers in future studies. On average participants needed a median of 83 minutes to complete the survey. The fastest was submitted after 12 minutes. Some participants took more time since the deadline to complete the project was set to one week.

4.1.1 Participants' Review Criteria

To provide insights into the security awareness and focus of freelancers, we evaluated the criteria which participants mentioned to have looked for in their code reviews. A detailed list of criteria mentioned by our participants is available in the Appendix E. We categorized the answers as follows:

Implementation: A total of six participants said that functionality was one criteria they looked for. Logic was mentioned by eight participants to be an important topic to look for in source code. One participant reported to have looked for maintainability and two mentioned performance and efficiency (PB5, NP4). A large number of participants (14/44) reported to also have checked for error handling.

Testing and bugs: NM3 said that quality assurance was one criteria to check for, while PB7 checked for unit tests and whether all scenarios are considered. NM1, NM5, and NP4 said they looked for bugs or errors in the code. A number of participants mentioned syntax to be a criterion to look for.

Standards and validation: NB1 and PB1 said that they checked whether code standards are met in the code. Three participants mentioned that code format was a criterion they checked for (PB2, NM4, PP5). Nine participants reported that they looked for the correct usage of get and set methods. NM1, NP2, NB8 mentioned the inspection of the model view controller architecture. Further, several participants reported to check imported packages and libraries. Some participants mentioned input validation and null checks (PM6) as criteria they checked the code for. Additionally, several freelancers reported to have checked for code style issues; e.g., camel case or naming conventions in the code. PB1 and PB2 wrote that they looked for duplicated code or unused code. Furthermore, four participants reported to assess the code complexity. Some participants said that they checked the code for readability and comments.

Security: Security in general was mentioned as a criterion by 10 participants. Eleven participants specifically included password storage security in their criteria. Data security was mentioned by 4 participants.

4.1.2 Found Password Storage Issue

Thirteen participants specifically mentioned in their reviews that secure password storage is an issue. Of these, only 2 were non-prompted. To solve the issue, PP8 suggested to use an external authentication service:

> "Depending on the application it may also be better in this case to simply use an external auth service such as that offered by google" (PP8).

Some prompted participants misinterpreted our prompting task description "Please ensure that the user password is stored securely" as password validation (NM3, PP4, PB8). For example, PB8 included secure password policies in the review but failed to detect the insecure password storage method:

> "The password must be at least 8 characters long. The password must have at least one uppercase and one lowercase letter. The password must have at least one digit. This needs to be updated" (PB8).

NM3 mentioned another password validation policy issue:

> "The most important part is the one you are not verifying the password what if it is equal to username. You should know that any person who is trying hit and try on the passwords will definitely first try to enter same username and password and he might be successful in your code and it's the worst part in security risks."

We also found that a number of participants used "password encryption" as a suggestion in their review, which is a concept not recommended for secure user password storage in a database. Furthermore, the prompted participant PB6 wrongly stated that the code snippets contained SQL and JAR injections, but did not mention insecure password storage as an issue. Finally, PM3 perceived the password storage implementation as "too complex" and asked in their review for more comments in the code snippet. A detailed list of code issues reported by our participants is available in the Appendix F.

4.1.3 Security Score

A number of participants commented on the password storage methods of the code snippets. Some participants explicitly said that the password storage functions were sufficient (NM4, NM5, PB8), others saw issues with the functions.

For example, a number of participants recommended using secure hashing functions like bcrypt (PM1, PM5), PBKDF2

(PM6), ARGON2 (PM1) or scrypt (PM5). In contrast, others recommended less secure functions such as MD5 (PP4), SHA-1 (PM4, PP4) or SHA-2 (PM1) (without mentioning iterations). It is noted that SHA-1 and SHA-2 can be secure when used with a key derivation function like PBKDF2, but we cannot assume that the participants mean this if they do not include it in their reviews. The code reviews of the 13 participants who identified user password storage security as an issue often lacked details. Thus, we only were able to grade 11 participants by using the security scale. Table 3 summarizes all the participants who found an issue with user password storage and their score according to the security scale. Only two participants specifically mentioned that a salt should be used.

4.1.4 Distraction Tasks

Some participants found the issues we introduced as distraction tasks. The logical mistake was found by 8 participants and the exception swallowing was mentioned by 12 participants.

NM5 falsely stated that error handling was done correctly. Out of all the 16 participants who found at least one distraction issue, 13 did not find the password storage issue, and 3 did (PM7, PB1, PM4). This might indicate that the distraction issues could have indeed distracted the participants from security. However, from the 13 participants who found the password storage issue, only 3 found at least another distraction issue.

4.1.5 Ready for Release?

We asked our participants to choose whether the code can 1) be released, 2) be released but the issues mentioned in the review need to be fixed for the next update, or 3) whether the code did not pass the review. Out of 44 participants, 2 said that the code can be released. Both found no security issues. Another 17 said that the code can be released but the issues should be fixed for the next update. 12 of these 17 participants did not find the security issue, so they referred to non-security related issues which needed to be fixed. The remaining 25 said the code did not pass the review and should not be released until the issues were fixed. 17 of these 25 participants, however, did not find the security issue with password-storage, so they have based their decision to not release the code on issues other than security. A detailed overview can be found in the Appendix D.

4.1.6 Participants' Definition of Hashing and Salting

After the review completion, we asked our participants to give the definition of hashing and salting passwords. We wanted to find out how many participants were able to correctly define hashing and salting of passwords. On average the participants took 8 minutes to answer this question (md: 5). We evaluated the responses and also tracked whether the participants left the tab inactive while answering the question. Out of 44 participants, 20 participants left the tab inactive for some time while 24 did not leave the tab to answer the question. The participants who left the tab inactive spent a median of 3 minutes outside the tab (mean: 6 minutes). This might indicate, that participants were searching the Web for the answer.

The majority of participants, 63% (28 of 44), gave a correct definition of hashing and salting for password storage. We checked whether the responses matched with definitions from the Internet indicating they were copied and pasted. We found that 8 participants copied the entire definition or parts of their definition from the Internet.

Table 3: Security score of participants who found the password storage issue

	Snippet	P	Score	Salt	Suggestion
NP1	Plaintext	n	0	-	"hashcode generation or convert in Hexa or other formats"
NP5	Plaintext	n	0	-	-
PB1	Base64	p	0	-	"an encoded format"
PB2	Base64	p	2	Y	-
PM1	MD5	p	3	-	"SHA2, Argon2, bcrypt"
PM4	MD5	p	1	-	"SHA-1"
PM5	MD5	p	3	-	"bcrypt, scrypt"
PM6	MD5	p	4	Y	"PBKDF2"
PM7	MD5	p	0	-	-
PP1	Plaintext	p	1	-	-
PP2	Plaintext	p	0	-	-
PP3	Plaintext	p	1	-	-
PP4	Plaintext	p	1	-	"SHA, MD5"

p = Prompted, **n** = Non-prompted

4.1.7 Security Responsibility

We asked our participants whether they felt responsible for end-users' security when writing or reviewing code. Figure 1 visualizes the responses. The "disagree" options 1,2,3 are summarized on the left. By contrast, the "agree" options 5,6,7 are summarized on the right. The percentages next to the bars are the sums of participants' ratings for each side. Neutral (4) is counted as an own point. For both tasks, the participants reported to strongly agree with the statements. However, after completing their reviews, we asked the participants whether they had ensured that the user password was stored securely. Out of 44 participants, 29 answered that they ensured that they had, which did not correspond to the evaluation of their

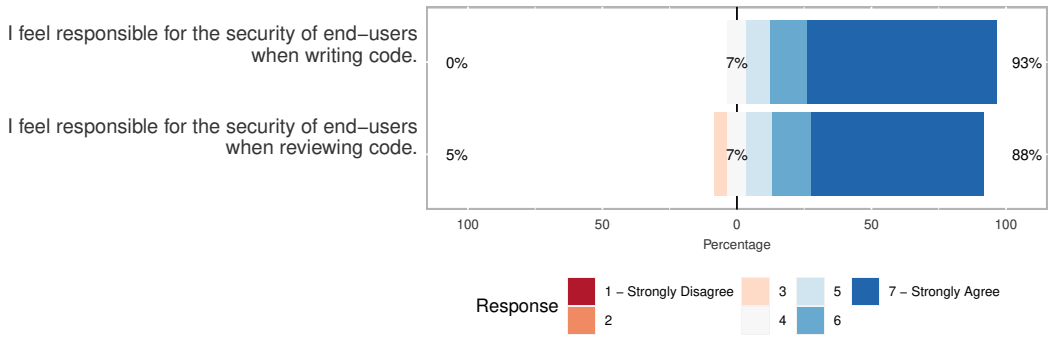


Figure 1: The responses on whether the participants feel responsible for security while code reviewing and writing.

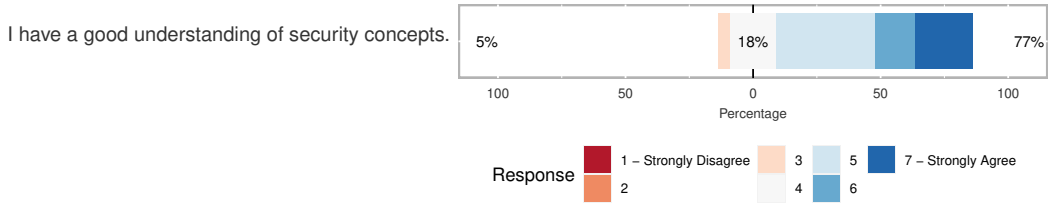


Figure 2: The responses on whether the participants reported to have a good understanding of security concepts.

reviews. In fact, only 13 participants correctly mentioned that insecure password storage was an issue in the snippets. This suggests a social desirability bias while answering survey questions, which was also reported by Naiakshina et al. [23] in their programming study with freelancers.

Additionally, we found an overconfident self-representation of the freelance participants. The degree of agreement to the statement around the freelancers' understanding of security concepts can be found in Figure 2. PP1, NB2, NM2, NP1, and PM6 specifically noted that security is very important in the optional feedback field. For example, NB2 noted:

> "Security is always important, developers put it in the background."

However, 3 participants explicitly noted in the optional feedback field that security is not always important, e.g., "for robotics control" (PP8). Further, PM4 stated:

> "In my experience, sometimes its not all about the security. Some occasions we have to provide hot fixes for urgent customers without thinking about the security. Yes security is an essential factor but it is not something that should be burden to a developer."

4.2 Quantitative Analysis

In this section we report the results of our exploratory quantitative analysis. We tested whether prompting, programming experience or different insecure code snippets had an effect on finding the password-storage security issue. We also tested whether prompting affected the word count of the code reviews and whether the self-reported time participants spend on security differs between programming and code-review tasks.

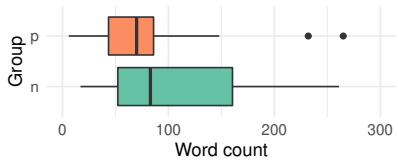


Figure 3: Word count within the reviews for the prompted and non-prompted group.
p: Prompted **n**: Non-prompted

4.2.1 Effect of Prompting on the Word Count

On average the reviews contained 100 words (md: 78 words) with the smallest review containing 6 words and the largest 443 words. We did not find a significant effect of prompting on the word count (Wilcoxon rank-sum, W = 300.5, p = 0.17). Figure 3 visualizes the word count in both groups. Both medians were in a similar range, however, the non-prompted group has a higher variable spread than the prompted group. The number of codes emerging from a code review did not necessarily rely on the word count in the review. Short reviews could cover different issues, while elaborate reviews

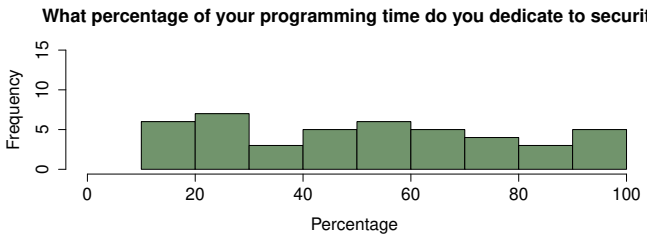


Figure 4: Percentage of *programming time* freelancers dedicate to security (self-reported)

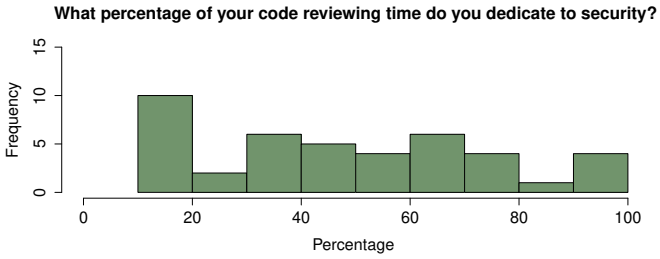


Figure 5: Percentage of *code reviewing time* freelancers dedicate to security (self-reported)

with more details could discuss only one minor issue.

4.2.2 Effect of Prompting on Finding the Password Storage Issue

We evaluated whether participants correctly indicated that there was an issue with password storage security in the snippet (*found password storage issue*). We excluded participants who wrote that the passwords should be "encrypted" if no hashing function was recommended. This might mislead the developer receiving the review to implement encryption instead of hashing the passwords. Naiakshina et al. [23] reported some encryption solutions, which shows that this might be a problem. Eleven prompted and two non-prompted participants correctly stated that password storage was not solved securely in the code snippet. Thus, prompting had a significant effect on finding the issue in the code snippet (FET: p = 0.008*, $cor - p$ = 0.02*, CI = [1.44, 89.85], OR = 8.28).

4.2.3 Effect of Experience on Finding the Password Storage Issue

In [13], Edmundson et al. did not find a significant effect of years of programming experience and whether the reviewers are more accurate or effective. We also investigated whether programming experience had an effect on whether participants found security issues with password storage. We counted how many of our 3 issues the participants were able to find (password storage, 2 distraction tasks). Similar to Edmundson et al., we did not find a significant correlation between the number of issues and the years of general experience (r = 0.05, p = 0.73). Further, we did not find a significant correlation between the number of issues and years of Java experience (r = 0.06, p = 0.72). We also did not find a correlation between Java experience and whether participants found the password storage issue (r = 0.06, p = 0.71).

4.2.4 Effect of Different Insecure Code Snippets on Finding the Password Storage Issue

All the three code snippets (plain text, Base64, MD5) showed an example for insecure user password storage in a database. We tested whether the different snippets had an effect on whether participants found an issue with secure password storage. For example, participants presented with an MD5 example might rather report an issue with password-storage security than participants presented with a code snippet where passwords were stored as plain text. We found, however, no significant effect within each subsample, neither the non-prompted (FET: p = 0.07, $cor - p$ = 0.14) nor the prompted group (FET: p = 0.26, $cor - p$ = 0.26).

4.2.5 Time for Security

While Figure 4 visualizes the distribution of percentages of *programming* time dedicated to security, Figure 5 summarizes the distribution of percentages dedicated to security during *code reviews*, according to our participants. The reported median percentage of time that the participants dedicated to security during code reviewing was 50 (mean: 51.64, min: 15, max: 100, sd: 26.54). The reported median percentage of time that the participants dedicated to programming was 55 (mean: 54.89 min: 10, max: 100, sd: 26.31). We did not find a significant difference between both reported time estimations using the sign-rank Wilcoxon test (V = 247, p = 0.12).

5 Discussion

RQ1: Developers' behavior in a security-critical code-reviewing task: Code reviewing is a technique applied at the end of the SDLC, used as one of the final steps by software developers to ensure programming code quality before software release. In comparison to programming code creators, developers take roles as programming code inspectors, which might increase their security awareness. Our study results showed, however, that this is not necessarily the case. It is alarming that almost half the participants wanted to release the insecure code snippets, although security issues with password storage can endanger millions of end-users' data. Even if participants indicated secure password storage as an issue, they often suggested poor techniques or weak hashing algorithms to improve the code. Such poor suggestions, however, might initiate the code

creator to revise the code without really improving security. It seems freelance developers need to be reminded of security during code reviewing, otherwise they might be too focused on other issues such as logical mistakes, conventions etc.

RQ2: Factors influencing developers' security awareness: We did not find an effect between which insecure programming code snippet the participant received and whether the participant reported issues with secure password storage. This is especially interesting, considering the fact, that they not only involved plain text password storage, but also Base64 encoding and even MD5 as a hashing function. Furthermore, we did not find an effect of programming experience on finding the insecure password-storage issue, which might indicate that more experience does not necessarily mean that the reviewers are more security aware or effective in finding security issues.

However, similar to the programming studies of Naiakshina et al. [22–24], we found that prompting for security in the reviewing task had an effect on finding the password storage issue in the code snippet. This means, that only if security requirements were mentioned in the task description, do participants consider security issues with password storage in their code reviews. Our results suggested that similar to programming, security needs to be part of the task during code reviewing as well. Therefore, we recommend to prompt for security when a code review is required.

RQ3: Developers' security responsibility in a code review: Our participants reported to spend half their programming and code reviewing time on security. We did not find a significant difference of the reported time spent on security between programming and code reviewing. Additionally, almost all the participants indicated to strongly agree with the statement that they feel responsible for security during programming and code reviewing and that they have a good understanding of security concepts. 66% of the participants also indicated to have ensured that the user passwords were stored securely after their code reviewing task. However, considering that only 13 of 44 (30%) participants reported a security issue with password-storage and the fact that almost all of them were prompted, this might indicate a social desirability bias in surveys. Our qualitative analysis showed that a number of participants had misconceptions and outdated knowledge of secure password storage. This might also suggest that APIs and libraries need to provide safe security defaults instead of requiring software developers to choose security mechanisms.

RQ4: Methodological implications: There is only limited knowledge of using code reviewing as a methodology for security studies with developers. While we provide insights into freelancers' behavior in code-reviewing tasks, we also wanted to explore which advantages, disadvantages and parallel insights a code-reviewing study can have in comparison to a programming study with developers. While we cannot conduct a direct comparison to the study of Naiakshina et al. [23] due to methodological differences, we still discuss the methodology of code reviewing for developer security studies by comparing the advantages, the disadvantages and some parallel insights of both the study types.

One disadvantage was the lack of certain information. We were not able to calculate the security scores of participants in such detail as Naiakshina et al. did. We could not find all information for the security scores in the reviews since participants simply did not mention them. Checking whether participants found the password storage issue was, however, still possible.

Moreover, we found that prompting had an effect on participants' solutions. This indicates that researchers investigating the security awareness of freelance developers might not need to hire them for longer programming tasks. Short and focused code reviews can offer similar results. With a median of 83 minutes to complete the survey, our participants required less time than Naiakshina et al.'s participants, who worked about 6-8 hours on the programming tasks.

Furthermore, code reviewing tasks can give indications to problems with code writing. Similar to Naiakshina et al., we were able to identify different issues developers experienced with password storage. For example, MD5 and encryption were often mentioned as adequate solutions to solve the password-storage issue. However, MD5 is an outdated hash function, which is not recommended any more for secure password storage. With encryption, participants might have referred to symmetric encryption [23], which is, as mentioned before, a discouraged practice for secure password storage. This suggests that code reviewing studies can offer valuable insights into participants' security behavior. We acknowledge though, that code reviewing is a different process to writing code and therefore it is not possible to prove which suggested solutions to the issues developers would really implement.

Similar to Naiakshina et al.'s password-storage study with students, we found that "security knowledge does not guarantee secure software" [24]. Although only 30% of our participants indicated that the user passwords were stored insecurely, 63% were able to provide a correct definition for hashing and salting after their code reviewing task. We have to note, however, that we had only limited possibilities to prove that their definitions were not simply copied and pasted from the Web. Its seemed that 8 participants copied the entire definition or parts of their definition from the Internet, which indicated that knowledge questions should be treated with caution in surveys.

To sum up, we found that security prompting had a significant effect regardless of whether participants completed a programming or a code-reviewing task on password storage. Additionally, we were able to identify participants' misconcep-

tions and outdated knowledge about secure password-storage and which criteria they believe are important in programming code. One disadvantage, however, was that we were limited in the comparison of the participants' security scores, which Naiakshina et al. introduced in their programming study. In our study, the code reviews did not offer enough details to calculate them. Still, code reviewing tasks can help investigate programming knowledge and decrease the time developers need to spend on a task. Participants needed less time to complete the study compared to the studies of Naiakshina et al. while still finding similar results with regard to security awareness and security prompting.

While we do not argue to replace programming tasks with code-reviewing tasks in security developer studies, funding is often limited within academia and smaller tasks yielding similar effects could enable more future research with developers. Therefore, we encourage the community to conduct further research into this line of work.

6 Conclusion

We conducted an online code reviewing study with 44 freelance developers showing each of them an insecure password storage code snippet. We investigated how participants behave in a code-reviewing study by considering which criteria they base their reviews on, whether they would find the security issue and most importantly, whether they would release the insecure code snippets. Additionally, we explored different factors, which might influence their behavior. For example, we explored the effect of prompting for security in the task on whether participants reported password storage security issues within their code reviews. We also explored whether participants feel responsible for and how much time they dedicate to security. Finally, we discussed the methodological implications of a code reviewing study for developer security studies.

Not even one third of our participants reported the security issue with password storage. Almost all the participants who reported an issue were prompted for security. Thus, prompting had a significant effect on participants' behavior. Still, almost half the participants wanted to release the code as it is, which is alarming since insecure password-storage is a major issue endangering millions of users. Finally, our findings suggest that code reviewing studies could be an interesting approach for conducting security developer studies.

For future work we recommend testing a hybrid between a code reviewing and a code writing developer study: a participant could receive functional insecure code and be asked to write a review and if necessary to correct the issues within the code. This could combine the advantages of both the methodologies. However, it might also increase the time of solving the study for the participants again.

Acknowledgments

This work was partially funded by the ERC Grant 678341: Frontiers of Usable Security.

References

[1] Fiverr.com. Accessed: September 2020.

[2] Y. Acar, M. Backes, S. Fahl, S. Garfinkel, D. Kim, M. L. Mazurek, and C. Stransky. Comparing the Usability of Cryptographic APIs. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 154–171, May 2017.

[3] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L Mazurek, and Christian Stransky. You Get Where You're Looking For: The Impact Of Information Sources on Code Security. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 289–305. IEEE, 2016.

[4] Yasemin Acar, Sascha Fahl, and Michelle L Mazurek. You are Not Your Developer, Either: A Research Agenda for Usable Security and Privacy Research Beyond End Users. In *Cybersecurity Development (SecDev), IEEE*, pages 3–8, Piscataway, NJ, USA, 2016. IEEE, IEEE Press.

[5] Yasemin Acar, Christian Stransky, Dominik Wermke, Michelle L Mazurek, and Sascha Fahl. Security Developer Studies with GitHub Users: Exploring a Convenience Sample. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 81–95, 2017.

[6] Hala Assal and Sonia Chiasson. Security in the Software Development Lifecycle. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 281–296, 2018.

[7] Hala Assal and Sonia Chiasson. 'Think Secure from the Beginning': A Survey with Software Developers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 289:1–289:13, New York, NY, USA, 2019. ACM.

[8] Alberto Bacchelli and Christian Bird. Expectations, Outcomes, and Challenges of Modern Code Review. In *Proceedings of the 2013 International Conference on Software Engineering*, ICSE '13, page 712–721. IEEE Press, 2013.

[9] Tobias Baum, Olga Liskin, Kai Niklas, and Kurt Schneider. Factors Influencing Code Review Processes in Industry. In *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*, FSE 2016, page 85–96. Association for Computing Machinery, 2016.

[10] Deanna D Caputo, Shari Lawrence Pfleeger, M Angela Sasse, Paul Ammann, Jeff Offutt, and Lin Deng. Barriers to Usable Security? Three Organizational Case Studies. *IEEE Security & Privacy*, 14(5):22–32, 2016.

[11] Anastasia Danilova, Alena Naiakshina, and Matthew Smith. One Size Does Not Fit All: A Grounded Theory and Online Survey Study of Developer Preferences for Security Warning Types. In *Proceedings of the 42nd International Conference on Software Engineering (ICSE'20)*, 2020.

[12] Cleidson R. B. de Souza, David Redmiles, Li-Te Cheng, David Millen, and John Patterson. Sometimes You Need to See Through Walls: A Field Study of Application Programming Interfaces. In *Proceedings of the 2004 ACM Conference on Computer Supported Cooperative Work*, CSCW '04, pages 63–71, New York, NY, USA, 2004. ACM.

[13] Anne Edmundson, Brian Holtkamp, Emanuel Rivera, Matthew Finifter, Adrian Mettler, and David Wagner. "An Empirical Study on the Effectiveness of Security Code Review". In Jan Jürjens, Benjamin Livshits, and Riccardo Scandariato, editors, *Engineering Secure Software and Systems*, pages 197–212, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[14] Satu Elo and Helvi Kyngäs. The qualitative content analysis process. *Journal of advanced nursing*, 62(1):107–115, 2008.

[15] Andy Field, Jeremy Miles, and Zoë Field. *Discovering statistics using R*. Sage publications, 2012.

[16] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.

[17] Peter Leo Gorski, Luigi Lo Iacono, Dominik Wermke, Christian Stransky, Sebastian Möller, Yasemin Acar, and Sascha Fahl. Developers Deserve Security Warnings, Too: On the Effect of Integrated Security Advice on Cryptographic API Misuse. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 265–281, 2018.

[18] Julie M Haney, Mary Theofanos, Yasemin Acar, and Sandra Spickard Prettyman. "We make it a big deal in the company": Security Mindsets in Organizations that Develop Cryptographic Products. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 357–373, 2018.

[19] O. Kononenko, O. Baysal, L. Guerrouj, Y. Cao, and M. W. Godfrey. Investigating code review quality: Do people and participation matter? In *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 111–120, 2015.

[20] Katharina Krombholz, Wilfried Mayer, Martin Schmiedecker, and Edgar Weippl. "I Have No Idea What I'm Doing" - On the Usability of Deploying HTTPS. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 1339–1356, Vancouver, BC, 2017. USENIX Association.

[21] Thomas D. LaToza, Gina Venolia, and Robert DeLine. Maintaining Mental Models: A Study of Developer Work Habits. In *Proceedings of the 28th International Conference on Software Engineering*, ICSE '06, pages 492–501, New York, NY, USA, 2006. ACM.

[22] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, and Matthew Smith. On Conducting Security Developer Studies with CS Students: Examining a Password-Storage Study with CS Students, Freelancers, and Company Developers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[23] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, Emanuel von Zezschwitz, and Matthew Smith. "If You Want, I Can Store the Encrypted Password": A Password-Storage Field Study with Freelance Developers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 140:1–140:12, New York, NY, USA, 2019. ACM.

[24] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, Marco Herzog, Sergej Dechand, and Matthew Smith. Why Do Developers Get Password Storage Wrong?: A Qualitative Usability Study. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pages 311–328, New York, NY, USA, 2017. ACM.

[25] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, and Matthew Smith. Deception Task Design in Developer Password Studies: Exploring a Student Sample. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 297–313, Baltimore, MD, August 2018. USENIX Association.

[26] Duc Cuong Nguyen, Dominik Wermke, Yasemin Acar, Michael Backes, Charles Weir, and Sascha Fahl. A Stitch in Time: Supporting Android Developers in Writing Secure Code. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1065–1077, 2017.

[27] Daniela Oliveira, Marissa Rosenthal, Nicole Morin, Kuo-Chuan Yeh, Justin Cappos, and Yanyan Zhuang. It's the

psychology stupid: how heuristics explain software vulnerabilities and how priming can illuminate developer's blind spots. In *Proceedings of the 30th Annual Computer Security Applications Conference*, pages 296–305, 2014.

[28] Marten Oltrogge, Yasemin Acar, Sergej Dechand, Matthew Smith, and Sascha Fahl. To Pin or Not to Pin—Helping App Developers Bullet Proof Their TLS Connections. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 239–254, 2015.

[29] Caitlin Sadowski, Emma Söderberg, Luke Church, Michal Sipko, and Alberto Bacchelli. Modern Code Review: A Case Study at Google. In *Proceedings of the 40th International Conference on Software Engineering: Software Engineering in Practice*, ICSE-SEIP '18, page 181–190. Association for Computing Machinery, 2018.

[30] Masha Sedova. Comparing Educational Approaches to Secure programming: Tool vs.TA. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, 2017.

[31] Dag IK Sjoberg, Bente Anda, Erik Arisholm, Tore Dyba, Magne Jorgensen, Amela Karahasanovic, Espen Frimann Koren, and Marek Vokác. Conducting realistic experiments in software engineering. In *Proceedings international symposium on empirical software engineering*, pages 17–26, Piscataway, NJ, USA, 2002. IEEE, IEEE Press.

[32] Davide Spadini, Gul Calikli, and Alberto Bacchelli. Primers or reminders?: The effects of existing review comments on code review. In *Proceedings of the 42nd International Conference on Software Engineering (ICSE '20)*, 2020.

[33] Jeffrey Stylos and Brad A Myers. The implications of method placement on api learnability. In *Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of software engineering*, pages 105–112. ACM, 2008.

[34] Mohammad Tahaei and Kami Vaniea. A survey on developer-centred security. *2019 IEEE European Symposium on Security and Privacy Workshops*, pages 129–138, 2019.

[35] David R Thomas. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation*, 27(2):237–246, 2006.

[36] Tyler W Thomas, Madiha Tabassum, Bill Chu, and Heather Lipford. Security During Application Development: An Application Security Expert Perspective. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.

[37] Sven Türpe, Laura Kocksch, and Andreas Poller. Penetration Tests a Turning Point in Security Practices? Organizational Challenges and Implications in a Software Development Team. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, 2016.

[38] Charles Weir, Awais Rashid, and James Noble. How to improve the security skills of mobile app developers? comparing and contrasting expert views. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, 2016.

[39] Chamila Wijayarathna and Nalin AG Arachchilage. Am I Responsible for End-User's Security? Baltimore, MD. USENIX Association.

[40] Chamila Wijayarathna and Nalin AG Arachchilage. Why Johnny Can't Store Passwords Securely? A Usability Evaluation of Bouncycastle Password Hashing. In *Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018*, pages 205–210, 2018.

[41] Jing Xie, Heather Lipford, and Bei-Tseng Chu. Evaluating interactive support for secure programming. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2707–2716, 2012.

[42] Khaled Yakdan, Sergej Dechand, Elmar Gerhards-Padilla, and Matthew Smith. Helping Johnny to Analyze Malware: A Usability-Optimized Decompiler and Malware Analysis User Study. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 158–177, San Jose, CA, USA, 2016. IEEE, IEEE.

A Survey

1. Thank you very much for your interest in our study. We are researchers from the University of Bonn. We are investigating the process of code reviewing for software developers. By taking part in our study you will help us understand how developers explore issues in source code written by other developers.

   Since we will be showing you code snippets please make sure to participate in the survey using your computer instead of a mobile device.

2. You and your colleagues are developing a social networking website to share pictures with family and friends. People need to register to this website in order to share their pictures.

   The front end has already been developed using the Model-view-controller (MVC) architectural pattern. A colleague of yours committed the following class and

you have been asked to do a code review for it, before it gets accepted and released.

The class implements a registration functionality in the web application: a user with the following attributes (email, name, gender, birth date, username and password) is stored in the database.

Please review the snippet as you would for a company. You get to decide if this code is deployed or if it goes back to the developer for improvement. In both cases please explain your decision. You can review the code manually or use tools if you prefer. (Prompting: Please ensure that the user password is stored securely.)
[CODE SNIPPET]

3. What did you check for? *[Text box]*
4. Please write your review for your colleague. Please suggest actions for all issues. *[Text box]*
5. Do you approve the release of the code?
   - The code has passed the review. The code can be released
   - The code can be released but the issues mentioned above need to be fixed for the next update.
   - Code did not pass review, please fix issues as mentioned above.
6. Did you check for security? *[Yes; No]*
7. Did you ensure that the user password was stored securely? *[Yes; No]*
8. Can you please explain what hashing and salting for passwords is? *[Text box]*
9. How did you review the code snippet? *[Manually; Using the following tools: [Text box]*
10. In the past I have reviewed code written by others. *[Yes; No]*
11. If Yes: How many times have you reviewed code written by others in the past year? *[Text box]*
12. What percentage of your code reviewing time do you dedicate to security? *[Text box]*
13. What percentage of your programming time do you dedicate to security? *[Text box]*
14. I have a good understanding of security concepts. *1 - Strongly Disagree - 7 Strongly Agree*
15. Please rate the following items: *1- Never - 7 Always*
    - How often do you ask for help when faced with security problems?
    - How often are you asked for help when others are faced with security problems?
16. I feel responsible for the security of end-users when writing code. *1 - Strongly Disagree - 7 Strongly Agree*
17. I feel responsible for the security of end-users when reviewing code. *1 - Strongly Disagree - 7 Strongly Agree*
18. Please enter your age: *[Text box]*
19. Please select your gender. *[Male; Female; Prefer not say; Prefer to self-describe: Text box]*
20. What is your current occupation? *[Freelance developer; Industry developer; Freelance tester; Industry tester; Academic researcher; Undergraduate student; Graduate student; Other:]*
21. What type(s) of software do you develop/test? (Multiple answers possible) *[Web applications; Mobile/App applications; Desktop applications; Embedded Software Engineering; Enterprise applications; Other (please specify):]*
22. In which country do you mainly work / study? *[Text box]*
23. How many years of experience do you have with software development in general?
24. How many years of experience do you have with Java development?
25. How many people work in your team? Please enter 1 if you work on your own.
26. Please select what is more important to you. *[Functionality - Security* (Slider between both, Middle: Equally important)

```
main{
 print(func("hello world"))
}

String func(String in){
   int x = len(in)
   String out = ""
   for(int i = x-1; i >= 0; i--){
       out.append(in[i])
   }
   return out
}
```

Figure 6: Test for software developing skills [11]

27. Please select the returned value of the pseudo code above [see Figure 6]:

- hello world hello world hello world hello world
- world hello
- hello world
- hello world 10
- HELLO WORLD
- dlrow olleh

28. As a non-profit academic institution we are interested in offering fair compensation for your participation in our research. How do you rate the payment of the study? *[Way too little; Too little; Just right; Too much; Way too much;]*

29. How many minutes did you actively work on this survey?

30. Thank you for taking part in our study! We really appreciate your time and effort. We hope our results will help improving security awareness in code reviewing. If you have any comments or suggestions, please leave them here and then please click on "Continue" to complete the survey.

B Play Book

During the study we conducted a play book to ensure all participants received the same information. When a seller contacted us, there were three cases: the offer is the correct amount, the offer is too expensive or the offer is too cheap.

- Hello! Thank you for your interest. We would be delighted to have you participate in our java code reviewing study. Do you have experience programming in Java? If you agree to proceed we would send you a link, from which you can then complete our online survey. We expect the survey to take no more than two hours. To complete the survey you would have a week. Would you like to proceed? Kind regards, XXX

- If the offer was not $50 we added the following question:

- If you would like to participate could you increase your payment requirement and send us a custom offer of $50?
  or
  We do however have a budget of $50 per participant. If you would like to participate could you send us a custom offer of $50?

- Once the participants had sent us a custom offer, they received the answer:
  Thank you! I will confirm your offer and then send you a link to the survey

- When you have completed the survey, we would appreciate it if you do not write any specific comments regarding the survey in your rating of us on Fiverr. As the study is currently ongoing this can lead to inconsistent results. Thank you for your understanding!

Below is a list of questions we were asked and our responses to them (P = Participant):

- P: I have very little programming experience in Java albeit.
  Us: We are looking for people with experience programming in Java. If you feel you fulfill this requirement you are welcome to take part.

- P: Is clicking on that link mean that I must start?
  Us: You should be able to continue where you left off, if you happen to want to continue the survey at a later point.

- P: I hope that the answers are to be in English?
  Us: Yes the survey is in English.

- P: I don't even know what is the problem and what is it about your research?
  Us: This is a Java code reviewing study. You will be required to complete a code review and then answer some questions.

- P: Why is that obligatory? (to get paid)
  Us: It is important for the study that each participant is treated the same.

- P: How many files / classes are and LOC (Lines of Codes) will be there in the code base? (roughly)
  Us: There are three files, two with roughly 100 lines of code and the third with 15.

- P: Do these two hours have to be without intervals?
  Us: You're welcome to take breaks as and when you need them.

- P: Just to get to know, do we need to do the survey straightaway for 2 hours or can we save the part that we have done and continue it later?
  Us: You don't need to complete the survey immediately. You can complete it at any point in the week after your offer is accepted.

- P: No personal information?
  Us: All data will be processed pseudonymously and stored anonymized after the study; there will be no identifying information published in any form.

- P: Seems a little sketchy to be honest, I'd like to make sure this is legit.
  Us: If you would like to participate could you send us a custom offer of $50? We would then accept your offer and send you the link, thus ensuring no risk for you.

- P: I wish to complete your online survey but Unfortunately paying you $50 is stopping me to participate.
  Us: You would be receiving the money.
- P: No I don't have any experience in Java.
  Us: Ok, thank you for your response!
- P: But how you know I take a survey and how you pay me?
  Us: You have sent us an offer of $50. I would confirm this offer and send you a link to our survey. Upon completion you will get paid.
- P: But I don't have any project if I don't deliver how it is possible to send money?
  Us: You have sent us an offer. As mentioned, I would confirm this offer, send you the link to our survey, which you would then complete. You would then confirm that you have delivered the service. We would then check that you have completed the survey. If this is the case, we will confirm completion and you will receive your payment.
- P: And what is the deadline? Is it limited by time?
  Us: You have a week to complete the survey.
- Participant claims to be finished, but the response is not submitted.
  Us: We have not received your response. Can you check that you have completed the survey?
- Participant mentioned word 'security' in review.
  Us: Thank you very much for your kind review. We have however noticed that you mentioned the word "security" in your review. As this study is ongoing, we would rather not have any comments regarding security on our profile. Is it possible you could change your review message? Kind regards, XXX
- P: They are asking for my review will you give me review otherwise I will mention that you haven't given me review after all work.
  Us: I have completed your review already.
- P: Please tell me and type here what review you want from me as seller.
  Us: Telling you what to review us is not in compliance with Fiverr's terms and conditions. We would appreciate it if you do not mention any specifics to the survey, but you are welcome to comment on the experience as a whole working with us.
- P: Do you have something new for me?
  Us: I'm afraid we don't have any more work for you at this time.
- P: What was the survey for? (After completion)
  Us: We are researchers working in the field of software usability. The survey is to be used to better understand how freelancers work and what benefits and disadvantages a code review has.
- By reapplication: Thank you for your interest in our survey, unfortunately we need new participants for the survey.
- Review: Very good communication, delivered on time. It was nice working with *name*!

C Code Snippets

Participants were shown at random one of three insecure code snippets. The code snippets for the study can be found here:

Plaintext

https://gist.github.com/u-cec/54e79635ec44234f8aa8ae4514d3d9e9

MD5

https://gist.github.com/u-cec/d25963ac45569962fca2291661f6e2f8

Base64

https://gist.github.com/u-cec/3fedf84f64918d9cafab61042cee8658

D Evaluation of Participants' Code Reviews

Table 4 shows an overview of the evaluation of participants' submissions.

E Participants' Review Criteria

Criteria mentioned by participants are summarized in Table 5.

F Found Password Storage Issue

Participants who reported issues with the code are summarized in Table 6.

Participant	Code Snippet	Prompted	Survey duration [minutes]*	Time for review [minutes]	Found password storage issue	Ready for release?	Review word count
NB1	Base64	n	31	18	0	✗	127
NB2	Base64	n	86	26	0	✗	56
NB3	Base64	n	24	4	0	✗	17
NB4	Base64	n	316	44	0	✓(!)	87
NB5	Base64	n	34	14	0	✗	128
NB6	Base64	n	29	12	0	✗	205
NB7	Base64	n	23	11	0	✗	63
NB8	Base64	n	36	27	0	✗	261
NM1	MD5	n	6119	30	0	✗	41
NM2	MD5	n	12	3	0	✓(!)	33
NM3	MD5	n	326	3	0	✗	443
NM4	MD5	n	62	38	0	✗	177
NM5	MD5	n	83	50	0	✓	117
NM6	MD5	n	254	195	0	✓(!)	79
NM7	MD5	n	85	49	0	✓(!)	40
NP1	Plaintext	n	83	37	1	✓(!)	74
NP2	Plaintext	n	151	118	0	✗	228
NP3	Plaintext	n	37	19	0	✗	73
NP4	Plaintext	n	3511	582	0	✗	171
NP5	Plaintext	n	40	29	1	✗	157
NP6	Plaintext	n	13	3	0	✓(!)	24
PB1	Base64	p	206	102	1	✗	265
PB2	Base64	p	934	46	1	✗	232
PB3	Base64	p	87	50	0	✓(!)	28
PB4	Base64	p	12	2	0	✓(!)	6
PB5	Base64	p	2407	132	0	✗	77
PB6	Base64	p	9045	38	0	✓(!)	82
PB7	Base64	p	68	41	0	✓(!)	30
PB8	Base64	p	198	122	0	✓(!)	70
PM1	MD5	p	25	14	1	✗	82
PM2	MD5	p	5798	1	0	✓	45
PM3	MD5	p	35	14	0	✗	45
PM4	MD5	p	129	105	1	✓(!)	148
PM5	MD5	p	6153	55	1	✓(!)	47
PM6	MD5	p	66	15	1	✗	84
PM7	MD5	p	39	15	1	✗	49
PP1	Plaintext	p	23	3	1	✗	25
PP2	Plaintext	p	2490	2	1	✓(!)	42
PP3	Plaintext	p	4444	67	1	✓(!)	87
PP4	Plaintext	p	25	12	1	✗	85
PP5	Plaintext	p	76	38	0	✗	107
PP6	Plaintext	p	198	14	0	✓(!)	89
PP7	Plaintext	p	53	5	0	✗	36
PP8	Plaintext	p	5910	21	0	✓(!)	68

Table 4: Evaluation of participants' code reviews

* Some participants started the survey and probably left it for some days since the deadline was to complete it within one week. ✗: Code did not pass review, please fix issues as mentioned above. ✓(!): The code can be released but the issues mentioned above need to be fixed for the next update. ✓: The code has passed the review. The code can be released. **Found password storage issue**: insecure password storage was mentioned as an issue in the review.

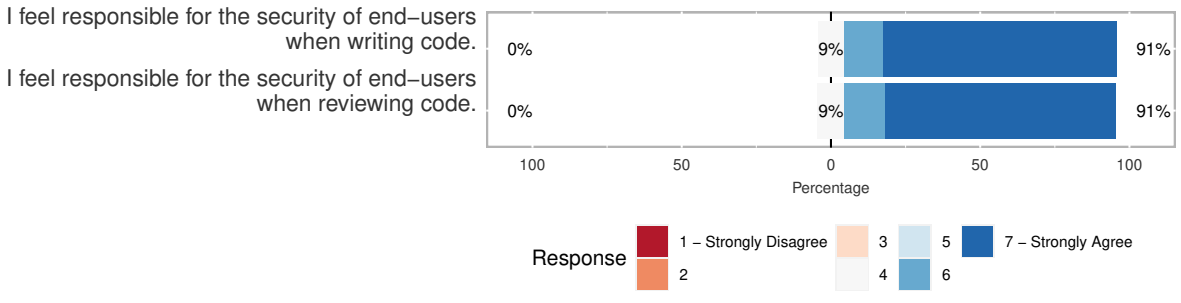


Figure 7: The responses on whether the participants feel responsible for security while code reviewing and writing (group: prompted)

	Criteria	Total Count	Participants	
			Prompted	Not prompted
Implementation	Functionality	6	PM4, PB6 NM5	NB6, NB1, NP1
	Logic	8	PM2, PM3, PP5, PB7	NB4, NP3, NB6, NM6
	Maintainability	1		NB2
	Performance / Efficiency	2	PB5	NP4
	Error Handling	14	PB1, PP3, PB3, PM3, PM5, PP5, PM7	NB1, NM2, NM4, NM5, NB6, NP4, NB8
Testing and Bugs	Quality Assurance	1		NM3
	Unit tests	1	PB7	
	Bugs/ Errors in the code	3		NM1, NM5, NP4
	Syntax	10	PM1, PB3, PM4, PP5 PB7, PP8, PM7	NM6, NP1, NP3
Standards and Validation	Code standards	2	PB1	NB1
	Code format	3	PB2, PP5	NM4
	Correct usage of get and set methods	9	PM1, PB3, PM5, PP5	NM2, NP1, NB3, NB4, NB5
	Model view controller architecture	3		NM1,NP2, NB8
	Imported Packages and Libraries	5	PP3, PB6	NP2, NB1, NB8
	Input validation	6	PM6, PP7, PP5	NB6, NM7, NM2,
	Null checks	1	PM6	
	Style issues	7	PB2, PM4, PP5, PP8	NB1, NP3, NP4, NB7
	Duplicated/ Unused code	2	PB1,PB2	
	Code complexity	4	PM4, PP5, PB5, PB7	
	Readability	5	PM4, PB5	NM1, NM4, NM5
	Comments	5	PB1, PM4, PB7	NB4, NP4
Security	Security	10	PB1, PP2, PP8, PM7, PM4, PP4	NM4, NP3, NP4, NP6
	Password Storage Security	11	PP1, PM1, PP2, PB2, PB6, PM6, PB7, PB8, PM7	NP1, NM4,
	Data Security	4	PB1, PP4, PM7	NM4

Table 5: All criteria mentioned by participants

	Found Issue	Total Count	Participants	
			Prompted	Not prompted
Found Password Storage Issue	Secure password storage	13	PP1, PM1, PB1, PP2, PB2, PP3, PM4, PP4, PM5,, PM6, PM7	NP1, NP5
	Password validation	3	PP4, PB8	NM3
	Password encryption	4	PP1, PP2, PP3	NP5
	SQL and JAR injections	1	PB6	
	Password storage to complex	1	PM3	
Security Score	Storage sufficient	3	PB8	NM4, NM5
	Function issue	5	PM1, PM4, PP4, PM5, PM6	
Distraction tasks	Logical Mistake	8	PM3, PP5, PP7	NM3, NM4, NB5, NP4, NB8
	Exception swallowing	12	PM4, PB1, PB3, PM3, PM7, PP5	NB1, NB2, NB6, NB8, NM4, NP3

Table 6: Issues found by participants

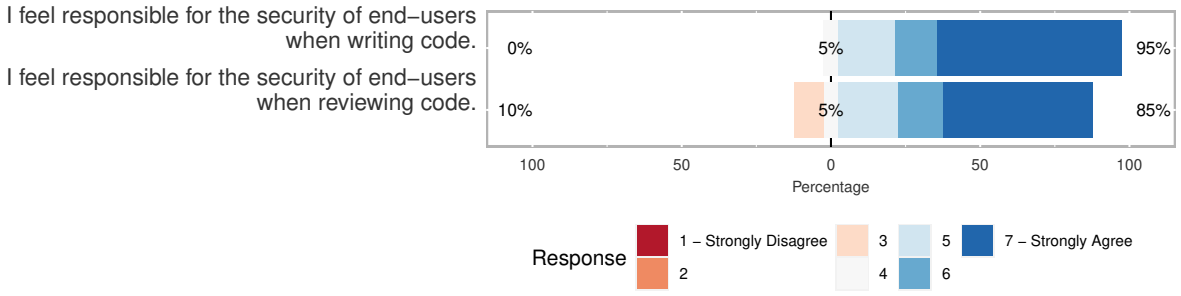


Figure 8: The responses on whether the participants feel responsible for security while code reviewing and writing (group: non-prompted)

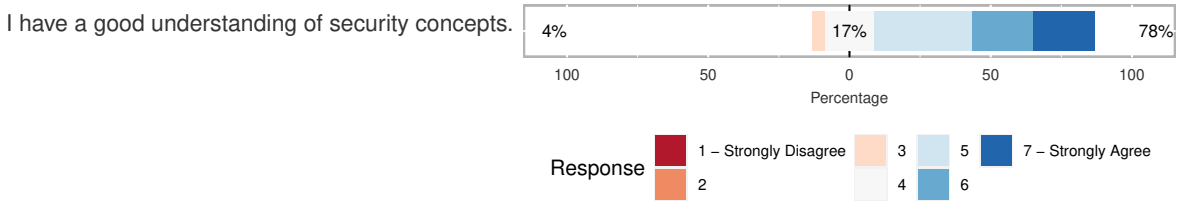


Figure 9: The responses on whether the participants reported to have a good understanding of security concepts (group: prompted)

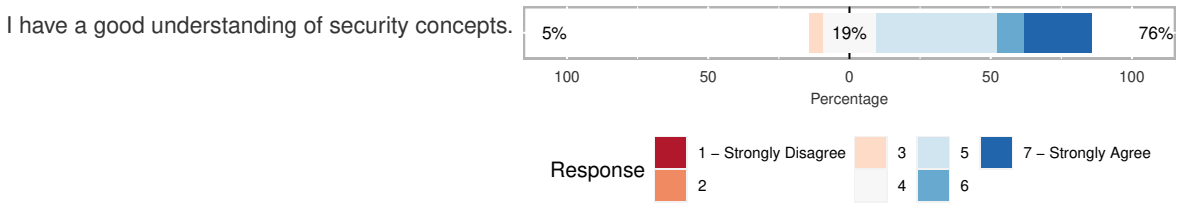


Figure 10: The responses on whether the participants reported to have a good understanding of security concepts (group: non-prompted)

"I have no idea what they're trying to accomplish:" Enthusiastic and Casual Signal Users' Understanding of Signal PINs

Daniel V. Bailey
Ruhr University Bochum

Philipp Markert
Ruhr University Bochum

Adam J. Aviv
The George Washington University

Abstract

We conducted an online study with $n = 235$ Signal users on their understanding and usage of PINs in Signal. In our study, we observe a split in PIN management and composition strategies between users who can explain the purpose of the Signal PINs (56 %; enthusiasts) and users who cannot (44 %; casual users). Encouraging adoption of PINs by Signal appears quite successful: only 14 % opted-out of setting a PIN entirely. Among those who did set a PIN, most enthusiasts had long, complex alphanumeric PINs generated by and saved in a password manager. Meanwhile more casual Signal users mostly relied on short numeric-only PINs. Our results suggest that better communication about the purpose of the Signal PIN could help more casual users understand the features PINs enable (such as that it is not simply a personal identification number). This communication could encourage a stronger security posture.

1 Introduction

Signal is an encrypted messaging application that is dedicated to preserving the privacy of its users and enacts features along those lines, such as not centrally storing users' contact lists, messages, or location histories unencrypted. Signal has historically relied only on users' telephone numbers for identification, authentication (via SMS), and contact discovery. Unfortunately, these methods are insufficient against attacks, including SIM-swapping [2, 18, 22]. In addition, these have some usability issues such as users who lose access to their telephone numbers also lose their Signal contact lists. Finally, they hamper additional features requiring additional metadata, like user profiles.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

To improve the app in terms of these shortcomings, Signal released two new features: *Secure Value Recovery* (SVR) [23] and *registration lock* [33]. Both features require the user to establish a PIN, which can be a sequence of numbers, like a traditional PIN, but also include letters and symbols. SVR uses the PIN to recover encrypted backups of contacts and settings stored on Signal servers. The registration lock aims to prevent anyone but the original user from creating a Signal account for a phone number without the associated PIN.

Signal's choice of naming the credential a "PIN" (as in, personal identification number) may not clearly indicate to the user the importance of the PIN in the Signal ecosystem. Unlike device or screen lock which is familiar to users, the in-app use of the Signal PIN is meant to achieve an app-specific purpose not satisfied by the device or operating system's features. A banking app for example might mostly be using in-app authentication to protect access to an OAuth token, while Signal has a different goal.

As Signal represents one of the first, large-scale usages of in-app PINs, in this paper we investigate to what extent do participants, both the security-/privacy-savvy and the average ones, understand the PIN feature and what effect does this have on their choice and usage? Additionally, we also investigate how participants react to Signal's PIN verification reminders that encourage users to not only select a complex PIN but regularly remind users to reenter it for verification. This feature may have been implemented because the PIN is not meant for daily use, but instead only needed in acute moments of setting up a new device with the Signal app. Finally, we examine the way participants select and compose their Signal PINs and the effect of their general understanding of the underlying Signal features to make these decisions. To this end, we consider the following research questions:

RQ1 Are participants aware of how and why in-app PINs are used in Signal?

RQ2 How effective are PIN reminders in assisting participants to remember PINs?

RQ3 How do participants choose and compose a PIN for Signal, and does their understanding of how these PINs are used affect that choice?

We surveyed Signal users (n = 235), asking about their understanding, usage of the Signal PIN feature, and response to Signal PIN verification. For example, we asked participants to explain the purpose of Signal PINs, in their own words. We additionally asked participants about the composition of their PIN (e.g., length, character set), if they reuse the PIN in other contexts (e.g., phone lock, in another messenger app), if they have opted out of selecting a PIN, and their response to periodic PIN verification.

We find that only 14 % (n = 33) of respondents opted out of setting a Signal PIN, and also we find a large disparity between the practices of participants who can explain the purpose of the in-app PIN authentication (who we term Signal *enthusiasts*; n = 132; 56 %) and those who cannot (dubbed *casual* Signal users; n = 103; 44 %).

Many enthusiasts set PINs because they thought it was required — initial communication from Signal indicated that it was, although it is not in current versions of the app. Many enthusiasts also specifically mentioned registration locking and cloud backups. Interestingly, when enthusiasts did not set a PIN, 44 % cited anti-cloud storage sentiments, indicating that they are aware of the features Signal PIN provides (e.g., cloud backups of profiles) but felt that this metadata storage did not sufficiently guard their privacy. Among casual users, 25 % set a PIN for generalized security reasons although they are not able to articulate those. Moreover, 13 % set a PIN simply because they were prompted by Signal or do not know why they actually set a PIN (16 %). If casual users did not set a PIN, they typically indicate that it was inconvenient (18 %) or they did not see the necessity (18 %). Their inaccurate understanding also affects this decision: 24 % state that they do not need an additional safeguard to secure access to their Signal app although the PIN is not used for this purpose.

Very few participants who set a PIN indicated that they had difficulty remembering their PIN; only 12 % said they *occasionally*, *frequently* or *very frequently* have difficulty remembering. When interacting with the periodic reminders to verify their PIN, 59 % confirm their PIN *frequently* or *very frequently*. Only 24 % of all participants confirm their PIN *rarely*, *very rarely*, or *never* when prompted, yet, here the behavior of enthusiasts and casuals diverges: 16 % of the latter tend to ignore the reminder prompt compared to 28 % of the enthusiasts. In addition, 45 or 24 % of the participants who currently use a PIN disabled these reminders. When asked why, 67 % of the enthusiasts mention that they use a password manager while casuals are mostly annoyed (42 %) or do not feel it is necessary to be reminded (33 %).

We also find that enthusiasts' PINs are more password-like, often containing numbers, letters and symbols. Compared to casuals, enthusiasts on average choose PINs with an additional 1.3 digits, 3.0 letters, and 1.3 special characters. Moreover, many participants, particularly enthusiasts, use a password manager to store their Signal PIN, which additionally increased the complexity of their PIN: password manager users selected PINs with an additional 2.1 digits, 5.3 letters, and 3.1 special characters compared to non-password manager users. A number of participants, both enthusiasts and casuals, noted the reuse of their Signal PIN in other contexts, apps, and as their screen lock, yet, 76 % of the participants who use a PIN within Signal said they do not reuse it.

In short, it appears Signal's core audience of privacy-conscious enthusiasts is using the PIN effectively, however, this roll-out may have been affected by inconsistent communication. Some earlier versions of the app made PIN creation a requirement. In addition, Signal PINs can contain letters and special characters. Weak Signal PIN choices can have consequences for those that choose secure PINs as secure communication requires both parties to be secure. We would recommend that Signal consider adding features to encourage better choices, like an improved blocklist, or even re-branding Signal PINs to more accurately depict their use, like "Account Recovery Passwords," which could help users apply the right context during selection and storage of this credential. Though our focus is on Signal, our results may inform communication strategies of other app developers, since account recovery and registration lock features are common in secure messaging.

All our findings were shared with the Signal developers.

2 Background

Signal is an open source app and service, developed and operated by the non-profit Signal Technology Foundation. Signal implements the underlying Signal protocol which includes forward secrecy [9, 38] and is used by other secure messaging clients, like WhatsApp [27] and Facebook Messenger's secret conversation feature [26]. Signal boasts more privacy consciousness in its design and implementation, eschewing linkages to an identity or collection of metadata, as compared to its competitors, like Telegram, WhatsApp, or Threema [43, 50]. Hereinafter, when we refer to *Signal,* we mean the app/service and not the protocol unless otherwise specified.

Given its focus on privacy, Signal historically relied on a user's mobile phone number as an identifier, reasoning that this system was already in place. This approach also makes migrating to a new device easier for users when using the same phone number, as long as the user's contacts were already backed up by other means. Other app settings, e.g., groups and blocked contacts, were formerly not backed up.

Additionally, receiving a valid SMS with a security code was sufficient to (re-)establish an account with Signal to send/receive encrypted messages. Unfortunately, phone numbers can be subject to SIM-swapping attacks [2, 18, 22], whereby an attacker is able to register an existing phone number with a new mobile SIM card, effectively stealing a user's account on Signal.

To address both backing up device settings and preventing account hijacking, Signal introduced two new features: *Secure Value Recovery* [23] and *registration lock* [33]. Both services require an additional authentication check, namely a PIN, and in the rest of this section, we describe Secure Value Recovery, registration lock, and how Signal rolled out PINs.

Secure Value Recovery *Secure Value Recovery* (SVR) enables encrypted backup and recovery of the Signal app settings, including contacts, profile, and group memberships. The backup data is encrypted and stored on Signal's servers. When a user migrates to a new device, the goal is to restore this data into the new app installation. As the decryption needs a key, the user has to choose, recall, and enter a PIN which is input to a key-derivation function. The resulting symmetric master key is used to further derive the backup encryption key.

Registration Lock The registration lock is an optional feature that binds the Signal PIN to the user's phone number. This way knowledge of the PIN is required as a second authentication factor in addition to the ability to receive an SMS with a one-time security code. This approach protects Signal from attacks like SIM swapping [2, 18, 22] where an attacker can obtain the SMS code.

To realize this functionality, the protocol uses the symmetric master key that is calculated as part of SVR, this time to derive a 32-byte registration lock hash. This value is used similarly to a password: it is sent to the server to authenticate the user. If the calculated registration-lock hash matches the one that is stored on the Signal server, the SMS code is sent. If not, the SMS code will not be sent and the registration of the phone number cannot be completed.

On the other hand, if an account needs to be migrated to a new device and the user does not know the PIN, setting up the account with the phone number is only possible after 7 days of inactivity. After this time span, the server's registration lock hash (of the PIN) expires and a new account can be created. However, the counter will be reset each time the client connects to the Signal server which happens when receiving or sending messages. Additionally, the iOS or Android apps make requests on a regular basis to keep the PIN hash alive even if the app itself is used infrequently.

Signal PINs Unlike PINs used to authenticate to gain access, e.g., unlocking your phone, the Signal PIN is used as a secondary authentication factor when moving an account from one device to another. A user does not need to enter the PIN to use Signal once it is installed on a particular device. However, Signal has a separate setting that locks the application from unauthorized access by forcing the user to verify their mobile phone's unlock authentication, e.g., the PIN used to unlock the device.

Also different than unlock authentication PINs, if a user forgets their Signal PIN while maintaining access to the Signal app, it can be reset without any repercussions as the current secure messaging keys can serve the purpose of authentication. After resetting the PIN, the SVR-encrypted backup can be re-encrypted and uploaded to Signal's servers, and the registration lock hash can be regenerated.

Communicating the purpose of the PIN to users, including all the features it does and does not support, is not a straightforward task. While Signal published an article explaining the technical details of SVR and registration lock [33], explaining it to all users remains a challenging task. Signal also originally required a user to establish a PIN, but later made that choice optional.

Finally, as the Signal PIN is only needed at acute moments, Signal employs periodic PIN reminders to help users memorize their PIN. These reminders to verify a PIN are spaced at regular intervals, starting at 12 hours, then 1 day, 3 days, 7 days, and every 14 days. Figure 1 shows the prompt that is shown to users for this purpose.

3 Related Work

The Signal PIN is used for authentication in the mobile setting, an area that has received a good deal of attention in the literature. For example, password usability on smartphones is studied along with shoulder-surfing resistance by Schaub et al. [40]. Aviv et al. consider the advantage gained by a guessing attacker as a result of screen smudges [3]. User choices for mobile unlock methods are investigated by Harbach et al. [16], who find that alphanumeric PINs are less popular than numeric PINs.

In our study, we note a number of casual users with limited comprehension, a theme also observed in other circumstances of secure messaging. Abu-Salma et al. [1] noted that security and privacy is not always a leading driver in the adoption of a secure messenger like Signal, but rather community pressure of wanting to be able to reach specific contacts. De Luca et al. [13] and Das et al. [10, 11, 12] come to a similar conclusion and show that the influence of social factors is not only limited to the adoption of messengers but security tools in general. Abu-Salma et al. [1] further note that many users have misconceptions about the security of messaging, e.g., they perceive SMS as secure for sensitive communication. Oesch et al. conduct a user study confirming user misconceptions and finding that group-chat users tend to manage security and privacy risks using non-technical means such as self-censorship and manually inspecting group membership [32].

In general, Signal and other secure messaging services often face the problem of explaining secure protocols, however, authentication ceremonies are challenging for users to understand [47, 48]. To address this issue, Wu et al. offered a redesign of the authentication ceremony that emphasizes comprehension [51]. Vaziripour et al. [46], on the other hand, suggested to partially automate the ceremony by using social media accounts. The Signal PIN is used for key derivation and is an example of a *usable encryption* scheme in the real

world. These have been previously studied by Ruoti et al. [39] who propose a secure email system and study varying levels of user transparency and automation. While Signal aims for automatic key management and automatic encryption, Ruoti et al. find that users had more trust in an approach that emphasized manual steps and therefore comprehension. While our research aims to understand how comprehension affects users' PIN practices, similar efforts to better communicate about this feature would likely help users.

As a user chosen secret, Signal PINs also relate to the choice of traditional PINs. Initial research on user choice of numeric PINs was done by Bonneau et al. [6] who found that dates are particularly prominent. Kim et al. also found dates to be common, as well as digits in sequence, like *1234* [20]. Signal's PIN in fact takes this advice and blocks digits in sequence. Wang et al. [49] derived numeric sequences from leaked password datasets, and Bonneau et al. [5] measured their guessability. Wang et al. found that PINs generally are easily guessable in online attacks (where an attacker only has a limited number of attempts or is rate-limited), and surprisingly 6-digit PINs more so than 4-digit PINs. This line of research was confirmed and extended in the context of mobile unlock PINs by Markert et al. [25]. Markert et al. also showed that a well-sized blocklist of PINs, when enforced, can significantly improve PIN-guessing resistance in an online setting.

Recently, Khan et al. [19] and Casimiro et al. [8] studied PIN reuse across different contexts. Both find that reuse is rampant, and that users tend to have a small set of PINs they use regularly. In our work we also find that certain kinds of PIN reuse is common for Signal PINs, such as for an ATM/Credit/Payment card. As Signal PINs are generally chosen and entered on mobile devices, users may be less inclined to choose hard-to-guess, full-fledged, alphanumeric passwords with special symbols. (Recall that a Signal PIN can have numbers, letters, and special symbols.) Melicher et al. studied user selection of passwords on mobile devices [28], finding that the limitations of the keyboard setting may lead to more easily guessable and weaker passwords.

In our work, we find that participants using a password manager are more likely to select strong Signal PINs. Unfortunately, in the mobile setting, users remain challenged in using password managers. Seiler-Hwag et al. investigated common password managers on smartphones [42], finding that all score poorly on standard usability metrics. Even when a password manager is adopted, using the password generation feature is not a given for all users. Pearman et al. [34] studied why users do (and do not) adopt a password manager and find that even those that do use a password manager may not use the password generation feature.

To the credit of the Signal team, they understood that the Signal PIN is unlike the case of mobile unlock authentication where a typical user unlocks the device multiple times per day. Instead, they realized an infrequently-used PIN is much more subject to being forgotten by the user. So they employ the well-known technique of *graduated interval recall* (also called *spaced repetition*). While the positive effects on recall rates have been shown in multiple studies [21, 29, 35, 44], including the memorability of passwords [4, 7, 17, 30, 31, 41], the usage of it in this context is novel. The deployment of Signal's periodic reminders to verify the PIN offers a real world example of the effectiveness of this strategy.

4 Method

We conducted a user study of $n = 235$ Signal users recruited to complete a survey about their understanding and strategies for managing their Signal PINs. In this section, we provide details of the survey, recruitment, limitations, and ethics.

4.1 Study Design

We recruited participants in two samples. The first sample was from Reddit, the Signal Community Forum, and snowballing; the second sample via Prolific. For participants completing the study on Prolific, we first used Prolific's built-in screening to only recruit participants who use Signal, and as this pool was still insufficient, we used a single screener question (Appendix A) as part of a two-part recruitment, where participants noted which messaging app they used. Those using Signal were invited to the main study. The entire survey is provided in Appendix B, and it took participants 7 minutes, on average, to complete.

1. *Informed Consent*: All participants were informed of the procedures of the survey and provided consent. The informed consent notified participants that they would be asked to complete a short survey that asks questions about how they select PINs and how they feel about Signal's implementation.
2. *Signal Usage*: Participants must indicate they are a Signal user answering the question: "Do you use Signal?"(**Q1**) All participants who responded in the affirmative continued with the survey.
3. *PIN Comprehension and Usage*: Participants were now prompted with the text: "PINs are a new feature provided by Signal. In your own words, please explain how PINs are used by Signal," (**Q4**) followed by "Did you set a Signal PIN?"(**Q5**) and why they did (**Q6a**) or why they did not (**Q6b**). Those who did not set a PIN skipped ahead to **Q25**.
   Those who did set a PIN were asked if the PIN was since disabled (**Q7**), and if so, why (**Q8**). We also asked participants who still had their PIN enabled if they have difficulty remembering their PIN (**Q9**), and what they would do if they forgot their PIN (**Q10**).
4. *PIN Reminders*: We then asked a series of questions (**Q11**-**Q14**) on Signal's periodic PIN reminders (cf. Figure 1), including if participants currently have the reminder set; for those who do, how frequently they verify the PIN when prompted; and if they disabled it, why.

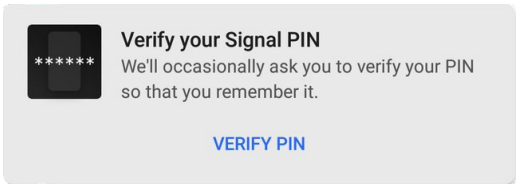


Figure 1: Prompt used by Signal to occasionally ask users to verify their PIN.

5. *PIN Reuse and Sharing*: Participants were asked to report if they reuse their Signal PIN in other contexts, such as mobile device unlock (**Q15**), ATM and other payment cards (**Q16**), and other mobile applications (**Q17**). In addition, participants were also asked if they have shared their PIN with friends or family (**Q18**). These questions were derived from related work on PIN usage [8, 19].
6. *PIN Selection and Composition*: The survey continued with a series of questions about PIN length and composition, as well as the perceived strength of the PIN (**Q19**-**Q24**).
7. *Other Messengers*: The survey continued by asking about the use of PINs in other messengers, including Facebook, Skype, Telegram, WeChat, and WhatsApp (**Q29**). We also asked if the Signal PIN is reused in any other messenger as well as the reasons for doing or not doing so (**Q30a/Q30b**).
8. *Demographics:* Finally, we asked about demographics (**D1**-**D5**), including age, gender, and IT background.

4.2 Recruitment & Demographics

We recruited a total of n = 235 participants. Of those 170 were recruited from Reddit, the Signal Community Forum, and snowballing, and 69 were recruited on Prolific. We posted to Reddit's r/SampleSize and r/Signal forums; and the Signal Community Forum. We decided against a fixed payment for these participants in favor of not collecting any personally identifiable information, e.g., an email address to offer a gift-certificate via a raffle, and thus these participants took the survey voluntarily without compensation.

We used Prolific's built-in custom prescreening filters, which allow researchers to post a study to participants that meet specific criteria, e.g., residing within the US. We applied the custom prescreening for Prolific members who indicated Signal is one of the "chat apps" they use regularly. We were able to recruit 69 participants this way, each paid GBP 1.50. To expand the Prolific pool, we also employed a custom screening survey to find other Signal users, recruiting 500 responses (paying GBP 0.15). Those who indicated that they used Signal were invited to the main study (paying GBP 1.50). We were able to recruit an additional 11 participants this way.

As shown in Table 1, the demographics of our sample is skewed toward a younger, more male-identifying, and more IT-oriented group. On the other hand, our participants reside in many different countries increasing the generality of our results. Of the 235 participants, Germany accounted for (68; 29 %), the USA for (61; 26 %); the UK for (24; 10 %). The rest of the world was the largest group with (82; 35 %). The actual demographics of the Signal community at large are unknown, so the skew towards a certain participant pool may reflect our recruiting strategy or may be influenced by the makeup of the underlying community. We observe that at 75 %, males make up the largest cohort. Similarly, at 64 %, those with IT-focused education or employment make up a majority of participants. In terms of education, bachelor's and master's

Table 1: Demographics of participants divided by subgroups.

	Enthusiasts		**Casuals**		**Total**	
	No.	**%**	**No.**	**%**	**No.**	**%**
Gender	132	56 %	103	44 %	235	100 %
Male	106	45 %	71	30 %	177	75 %
Female	13	6 %	24	10 %	37	16 %
Non-Binary	1	0 %	1	0 %	2	1 %
Other	1	0 %	0	0 %	1	0 %
Prefer not to say	11	5 %	7	3 %	18	8 %
Age	132	56 %	103	44 %	235	100 %
18–24	31	13 %	14	6 %	45	19 %
25–34	57	24 %	53	23 %	110	47 %
35–44	29	12 %	17	7 %	46	20 %
45–54	7	3 %	10	4 %	17	7 %
55–64	5	2 %	3	1 %	8	3 %
65–74	0	0 %	3	1 %	3	1 %
75 or older	1	0 %	0	0 %	1	0 %
Prefer not to say	2	1 %	3	1 %	5	2 %
Education	132	56 %	103	44 %	235	100 %
Some High Sch.	0	0 %	3	1 %	3	1 %
High School	31	13 %	12	5 %	43	18 %
Some College	0	0 %	0	0 %	0	0 %
Trade	0	0 %	4	2 %	4	2 %
Associate's	3	1 %	6	3 %	9	4 %
Bachelor's	35	15 %	32	14 %	67	29 %
Master's	38	16 %	25	11 %	63	27 %
Professional	9	4 %	3	1 %	12	5 %
Doctorate	10	4 %	12	5 %	22	9 %
Prefer not to say	6	3 %	6	3 %	12	5 %
Country	132	56 %	103	44 %	235	100 %
Germany	48	20 %	20	9 %	68	29 %
USA	25	11 %	36	15 %	61	26 %
United Kingdom	7	3 %	17	7 %	24	10 %
Other	52	22 %	30	13 %	82	35 %
Background	132	56 %	103	44 %	235	100 %
Technical	96	41 %	54	23 %	150	64 %
Non-Technical	33	14 %	44	19 %	77	33 %
Prefer not to say	3	1 %	5	2 %	8	3 %

groups combined account for 55 % of participants. Our group of enthusiasts is also male-dominated: self-identified males outnumber females more than 8:1. Finally, we note that among enthusiasts, the IT-focused group is substantially larger at 3:1, while the figures are more balanced for casuals: about 1.2:1. It is reasonable to surmise that an IT background makes one more likely to be a enthusiast — put another way, Signal's existing communication strategy about the Signal PIN appears to be more effective for those with an IT background.

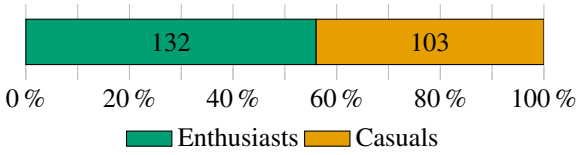


Figure 2: Classification of the participants based on the participants' ability to explain the usage of PINs in Signal (**Q4**).

4.3 Limitations

As this study took place online, it shares the usual limitations of many online studies, such as finding a representative recruitment. On the one hand, our sample may not be a representative sample of all Signal users. Though we did not explicitly sample enthusiasts and casuals separately, we found that comparatively more enthusiasts were recruited via Reddit and Signal Community Forum, which led us to perform additional sampling from Prolific.

As an online survey, this study necessarily relies on self-reported data. With regard to security and privacy user studies, Redmiles et al. [37] show online-survey responses generalize quite readily to the broader population. Additionally, we conducted extensive pilot testing among members of our research groups and trusted colleagues to identify any ambiguities in our survey questions.

Another limitation is that participants' responses may suffer from the well-known tendency toward providing socially-desirable answers [14, 24]. For example, it is possible that PIN reuse is more prevalent than our study suggests, or that people choose PINs that are shorter and have less-diverse composition. The same holds for questions where we asked participants about their own understanding, where they might have looked up answers on Signal's website. Despite this possibility, the answers provided appeared unique and participants provided many apt phrases to describe the situation. Additionally, we did not find responses that were directly cut and paste from Signal's website.

4.4 Ethics

The study was administered at an institution that does not have an Institutional Review Board (IRB), but we still followed all appropriate study procedures similar to studies that obtained IRB approval. For example, participants were informed about the nature of the study, participated voluntarily, and could opt-out at any time. Additionally, we conformed with the ethical principles laid out in the Menlo Report [45], e.g., we minimized any potential harm by not collecting any personally-identifiable information from our participants.

As described above, we completed two recruitments, one with paid and one with unpaid participants. Unpaid participants were recruited via Reddit, Signal Community Forum, and snowballing. We decided not to pay those participants as paying a comparatively small amount did not appear to withstand the harm that went along with collecting email addresses. Additionally, as this community tends to be more privacy-conscious, doing so might have depressed participation. Participants recruited via Prolific were paid GBP 1.50 for successfully completing the main survey, as this amount is in line with the recommended rewards on Prolific [36].

5 Results

In this section, we present the results of our study of $n = 235$ Signal users. For the structure of the section, we follow our three research questions: we start by analyzing the comprehension of the usage of PINs in Signal (**RQ1**), continue with user responses to the reminder feature (**RQ2**), and conclude with PIN selection and composition (**RQ3**).

For qualitative analysis, we had a primary coder code all the qualitative responses, producing an initial codebook. A secondary coder used that codebook to independently code the same responses, and afterward, the two coders met to resolve differences to produce a final codebook. The primary coder then used that final codebook to re-code the data. The codebook used for each qualitative question can found in Appendix D, Tables 3–11.

5.1 RQ1: Comprehension

As part of **RQ1**, we seek to understand Signal users' awareness and understanding of PINs and how they fit into the Signal ecosystem. To answer this question, we divide the participant pool by those that have or have not adopted a Signal PIN, and also by those that demonstrate understanding of how Signal uses the PIN.

Understanding Signal PINs After indicating if they are a Signal user (**Q1**–**Q3**), we first ask participants to describe how Signal PINs are used in their own words (**Q4**): *PINs are a new feature provided by Signal. In your own words, please explain how PINs are used by Signal.* These responses were coded by comprehension and accuracy; specifically, we seek to understand if the participants recognized that PINs are used for SVR and registration lock. Participants who accurately described the usage of Signal PINs were coded as *enthusiasts* ($n = 132$; 56 %), and those who could not describe Signal PIN usage were coded as *casual* Signal users ($n = 103$; 44 %).

We observed many different ways of capturing the main elements of how PINs are used by Signal. Many of the enthusiasts were even able to demonstrate a deep understanding, for example P10 said:

> *"It protects data like settings and group membership and signal [sic] contacts that will be stored on Signal's servers using SVR. Previously this was only stored locally on a user's device and was lost upon device reset or getting a new device unless a full backup was made on Android."*

Participant responses were assigned one or more codes based on the aspects correctly described. Overall among enthusiasts, the most popular codes were backup (65; 49 %), encryption (45; 34 %), contacts (31; 24 %), and registration (23; 17 %). Some also noted settings (8; 6 %), profile (4; 3 %) or groups (3; 2 %), which are also secured via a Signal PIN during backup, and a few specified key derivation (7; 5 %). Some also mentioned that PINs were part of a process for Signal to move away from using phone numbers for identity (6; 5 %). A handful of enthusiasts also expressed anti-cloud sentiments when asked about Signal PINs (2; 2 %), suggesting that they understood that the PINs play a role in the encrypted cloud backup functionality of SVR, and that they are opposed to that design direction.

For the casual users, a majority (57; 55 %) provided non-answers, or answers that do not indicate any understanding of the way the Signal PIN is used. The answer of P47 accurately summarized the reasoning we observed for many casual users:

> *"I don't understand their purpose very well. I thought that they might be using the PIN system to verify the identity of the person using signal (if for instance someone unauthorized gained access to the phone), but the way that pin entry is optionally offered every few weeks doesn't align with such a purpose. as such, I have no idea what they're trying to accomplish."*

As the majority of casual users didn't know or provided non-answers, there are many other examples to choose from, including "I initially thought it was used as a local PIN to unlock the app on my phone. It doesn't do that so I have no idea how it works," from P62. Additionally, many casual users falsely associated PINs with securing messages (21; 20 %) although messages are not part of the backed-up data and are not protected by the PIN, as explained by P183: "Keep your messages on Signal encrypted via use of the PIN."

An equal number felt that the PIN locks the Signal app (21; 20 %), while in fact that functionality is called Signal Screen Lock and is not related to the Signal PIN — for that feature, Signal simply re-uses the device's existing PIN, biometric, or other authentication scheme. An example of this response is from P37: "Protect application from opening from an unlocked phone." Similar responses show this is a common misconception: "Pins are used to prevent unauthorized access to the app" from P227. Some individual participants also mentioned security as a general topic, without further describing it (2; 2 %), or associated the PIN with inconvenience (1; 1 %).

Why did participants set a PIN? In addition to knowing if participants understand the usage of the PIN, we also want to analyze how many actually set a PIN in their Signal app. In total, 202 or 86 % of all 235 participants adopted a PIN. If we further divide those 202 participants based on their understanding, we see that more enthusiasts (116; 57 %) than casuals (86; 43 %) set a PIN.

To get a deeper understanding, **Q6a** asked participants to explain their decision. By far the most popular reason, equally distributed among enthusiasts and casuals, is *security*: 48 or 24 % mentioned it in their answer. Once again, we find that enthusiasts display a detailed, in-depth understanding, exemplified by P14:

> *"I want to be able to use secondary identifier once it becomes available and not to lose my contacts that are not in my phone's contacts list. I also want to be secure against SIM-swap attacks."*

This code is followed by participants mentioning that they were required to set a PIN (33; 16 %). Among enthusiasts, we observed 25 that mentioned it was required (or 22 %). P164 said "I had absolutely no choice if I wanted to continue to use Signal. Eventually, the box asking you to create a PIN kept you from opening any of your messages until you did what it wanted."

This response may reflect the changing nature of the PIN requirement. Initially, it was required and then in a subsequent version, merely encouraged. The enthusiast-casual split here suggests perhaps more enthusiasts were early adopters of the Signal PIN. Another theme, of setting a PIN due to annoyance (12; 11 %) may also reflect this changing communication strategy for Signal PINs. See for example Figure 8, showing the initial prompt used by Signal to ask users to create a PIN; the prompt has subsequently been updated to 9, current as of this writing. Observe the communication is also different when a user wishes to change their PIN as shown in Figure 10, again current as of this writing.

Enthusiasts also regularly noted registration lock as a reason to set a PIN (14; 12 %). P3 said "The PIN stop [sic] others from registering as me, and also protects access to my account details (profile, settings, contacts) if my device is misplaced."

Casual Signal users noted *security* most frequently (26; 29 %), but did so in a more general way as seen in this quote from P141: "for security and for reassurance if device gets stolen." Additional codes include *don't know* (16; 18 %) and *prompted* (13; 15 %), suggesting that many casual users selected a PIN simply because they were prompted to do so and had no other underlying motivations. For example, P155 responded "I trusted the app and just did it when prompted."

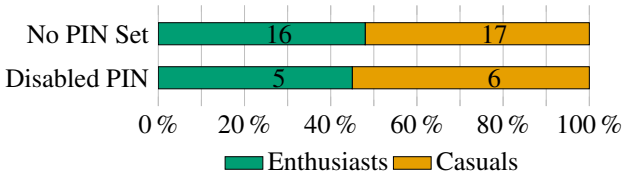


Figure 3: Classification of the participants who disabled or did not set a Signal PIN.

Why did participants not set a PIN? A total of 33 (14 %) participants chose not to set a PIN (see Figure 3). A roughly equal number of enthusiasts and casual Signal users did not set a PIN: 16 enthusiasts (12 %) did not set a PIN and 17 (17 %) casual Signal users did not set a PIN. A χ^2 test revealed no significant differences between the groups.

When these n = 33 participants described why they did not set a PIN (**Q6b**), there were a number of differences. Both casual (3; 18 %) and enthusiasts (4; 25 %) described PINs as inconvenient, but casual users were more likely to note that either they do not need a Signal PIN (3; 18 %) or that their phone lock provided security (4; 24%). For example P227 noted that their "…phone is always locked" and "Additional authentication seems unnecessary."

Enthusiasts expressed distrust as a reason for not setting a PIN. Either this distrust is in the security of PINs for key derivation and management (3; 19 %), or they distrust cloud storage (7; 44 %). Distrust of cloud storage stems from privacy concerns with the SVR feature that backs up contacts and settings. P216, for example stated, that they "had no desire to have any contact data uploaded," and P207 said "i [sic] do not want to store personal information in the cloud."

Why do participants disable PINs? On top of the 33 users who declined to set a PIN, a total of 11 (5 %) set a PIN and then later disabled it: 5 (45 %) enthusiasts and 6 (55 %) casual users, as shown in Figure 3. When asked to explain why they disabled their PIN (**Q8**), participants mentioned that the PINs were annoying (4; 36 %) or inconvenient (2; 18 %), which may be related to the periodic verification reminders. P212 explicitly mentioned the "verification overhead." Anti-cloud hesitation to store data on Signal's servers led (3; 27 %) participants to disable their PIN: "Don't want my data stored on their server" (P193). We also observed (2; 18 %) participants who simply stated that they "do not need it" (P206).

RQ1 Results Summary Signal users in our sample break down into two groups: enthusiasts who were aware of the features Signal PINs enabled, and more casual Signal users who were unable to describe how PINs are used within Signal. In both groups, though, setting a Signal PIN was highly prevalent. Only 33 of the 235 respondents chose not to set a PIN. Among enthusiasts, their choice to not set a PIN stemmed from either distrust in the key-derivation process or hesitancy to store information in the cloud generally. Casual users did not set a PIN because of inconvenience or a false belief that other authentication mechanisms, like locking their phone, provided adequate protection. When participants disabled their Signal PIN, inconvenience or annoyance were often cited, sometimes referring specifically to the periodic reminders.

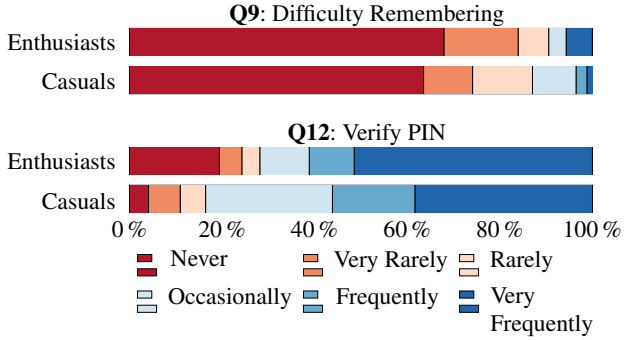


Figure 4: PIN memorability and verification.

5.2 RQ2: PIN Recall and Reminders

In this section, to address **RQ2**, we consider how participants remember their PINs and their reactions to the periodic PIN verification reminders. Throughout this section we consider the n = 191 participants who still have their PIN enabled, and not the 11 participants who since disabled their PIN.

Forgetting PINs We asked the (n = 191) participants who still use a Signal PIN in **Q9** if they encountered difficulty in remembering their PIN. Overwhelmingly, 89 % of participants (n = 170) indicated that they *never*, *very rarely*, or *rarely* have difficulty remembering their PIN (see Figure 4; *top*). We compared the response to this question from enthusiasts (n = 106) and casual (n = 85) Signal users who still had their PIN enabled, and we found no statistical differences.

We asked participants in **Q10** what they would do if they forgot their Signal PIN. (Note that the PIN is not required to use Signal for messaging, and can be reset at any time in the settings menu.) Many enthusiasts noted that their PIN was stored in their password manager (45; 42 %), and they would simply look it up. Fewer casual participants mentioned a password manager (12; 15 %). A number of participants did not know what to do (27; 25 % enthusiasts and 33; 40 % casuals), while a few casuals suggested they would contact Signal (4; 5 %) and two enthusiasts said they would reinstall the app (2; 2 %). Others believed that their Signal account is now unrecoverable (2; 2 % enthusiasts and 3; 4 % casuals); some would create a new account (4; 4 % enthusiasts and 5; 6 %) casuals. A handful (2; 2 % enthusiasts and 4; 5 % casuals) denied that they would forget stating "It is a PIN I use for my bank cards" (P145), for example. A small number of participants noted that they would wait (8; 7 % enthusiasts and 4; 5 % casuals), aware that the registration lock expires after 7 days of inactivity.

Periodic Verification Perhaps recognizing that Signal PINs are only truly required when transferring a Signal account to a new device, Signal decided to employ *graduated interval recall* [35] (or, *spaced repetition*) that regularly prompted participants to verify their PIN when opening the Signal app. An example of such a reminder is found in Figure 1. To our knowledge, Signal is the first mainstream app to implement such a feature.

We first asked participants if they were aware of the PIN verification reminders (**Q11**). Most participants (n = 176; 92 %) indicated that they were aware, and a follow up question (**Q13**) asked if they have since disabled the reminders. Seventy-four percent (n = 131) of participants have the periodic PIN verification enabled, and many still verify their PIN when prompted. Seventy-six percent (n = 135) of participants either *occasionally*, *frequently*, or *very frequently* verify their PIN when prompted. When dividing this data by enthusiasts and casual Signal users (see Figure 4; *bottom*), we did not observe significant differences between frequency of PIN verification using a Mann-Whitney U test.

The remaining 23 % (n = 45) disabled the PIN reminders. These 45 participants were asked why they disabled the reminders (**Q14**): (23; 51 %) mentioned doing so because they use a password manager. P63 said "I don't remember my PIN, it's stored in my password manager, frankly, I don't even want to remember it." Ten (22 %) said there was no need or their PIN was already memorized, and a further (11; 24 %) found the reminders annoying. These figures suggest that the periodic reminders are generally viewed as beneficial, or at least not substantially invasive enough to warrant disabling them. As we rely on self-reported data, we do not independently verify PIN recall rates.

Password Manager Usage We found a large amount of password manager (PM) usage in our study. These reports were entirely unprompted as PMs were not mentioned in any survey material. Thirty-one percent (n = 62) indicated that they use a PM in response to questions regarding either what they would do if they forget their PIN **Q10** or how they select their PIN **Q20**. As we did not explicitly ask about PM usage, the true number of PM users might be higher.

More striking is the combination of the classification of enthusiasts and casual participants combined with that of PMs: (52; 83 %) of the 62 participants who said they use a PM were enthusiasts. Or, 50 % of the 103 enthusiasts who have a PIN enabled use a PM. Only (10; 14 %) of the 73 casual Signal users using a PIN mentioned PMs as a mechanism to either select or recall their PIN. Put another way, participants who mentioned a PM were overwhelmingly enthusiasts.

RQ2 Results Summary Participants indicated that they have little difficulty remembering their PIN, many stating that this is a PIN they use all the time and thus would *never* forget it. A large number of participants, notably half of PIN-using enthusiasts, use password managers to both select and recall their Signal PIN, and are thus, not concerned with forgetting their PIN. Reactions to Signal's periodic PIN verification requests were more mixed, but overwhelmingly participants verified their PIN when prompted. Roughly a quarter of participants disabled periodic PIN verification; most did so because they use a password manager. Others stated that the PIN was already memorized, so there was no need for the reminders, and some simply found the reminders annoying. Overall, since 76 % of participants reported verifying their PIN when prompted, we conclude graduated interval recall used for Signal PIN verification is generally embraced by users, though the effectiveness of this intervention is obviously an area that deserves future work.

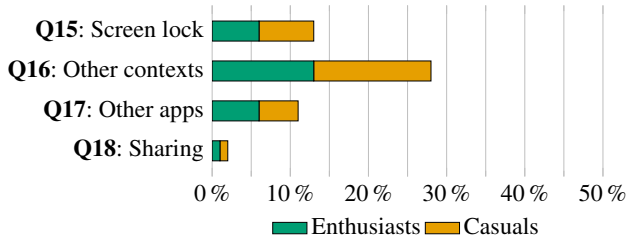


Figure 5: Frequency of PIN reuse and sharing.

5.3 RQ3: PIN Reuse and Composition

In this section, we explore selection strategies of Signal PINs by asking participants if they reuse their Signal PIN in other contexts; the composition of their Signal PIN with respect to numbers, digits, and special symbols; and the perceived security of their Signal PIN in comparison to other PINs they use.

PIN Reuse To explore the many ways in which PINs are reused, we adopted questions from Khan et al. [19] and Casimiro et al. [8] regarding PIN usage, more broadly. The responses of n = 191 participants using a Signal PIN are found in Figure 5, broken down by enthusiasts and casual users.

First, as a mobile application, we asked participants if they used their smartphone unlock PIN as their Signal PIN (**Q15**). Thirteen percent (n = 26) did so, composed of 12 enthusiasts and 14 casual users. In **Q16**, we asked if they used the Signal PIN in other contexts, ranging from ATM/Credit/Payment cards, to garage door codes, gaming consoles, and voice mail. (Refer to Appendix B for the full list, derived from Khan et al. and Casimiro et al.) Twenty-eight percent (n = 53) of participants use their PIN in another context, consisting of (25; 43 %) enthusiasts and (28; 53 %) casual users. Among those who reused, casual users did so more often: 1.39 times on average, compared to enthusiasts who did so 1.24 times. The most common context of PIN reuse overall was for ATM/credit/payment cards where (17; 32 %) of 53 participants reused a PIN. Participants also mentioned laptop/PC authentication (13; 24 %) and other online accounts (11; 21 %).

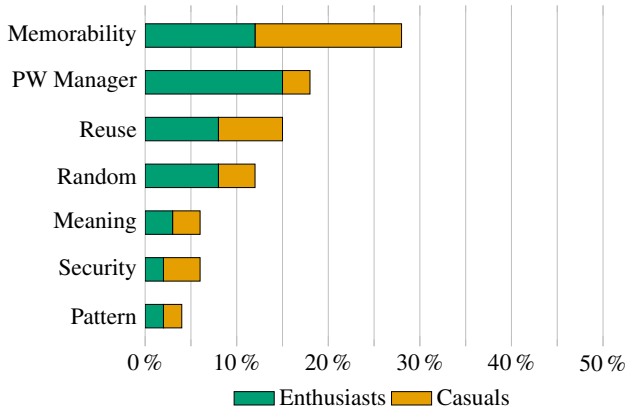


Figure 6: Most popular codes assigned to the answers of **Q20**: What was your primary strategy in selecting your Signal PIN?

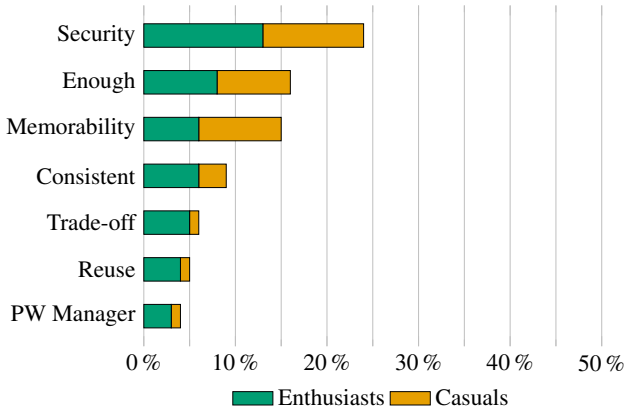


Figure 7: Most popular codes assigned to the answers of **Q22**: Why did you choose a PIN with this security level?

We also asked if participants reuse PINs in other mobile applications (**Q17**): (21; 11 %) reported they did, and of those, 12 were enthusiasts and 9 were casual users. Most commonly, the other app was WhatsApp (n = 6); WhatsApp implements the Signal Protocol. Other common mobile apps where this PIN was reused were banking apps (n = 5). In **Q25**–**Q28**, we asked participants if they use other messenger services, such as Facebook messenger, Telegram, and WhatsApp: (183; 95 %) did. We also asked if they set a PIN in these services and found (49; 26 %) did.

Finally, we asked if participants share their PIN with friends and family: this was rare. Only 3 participants did so, suggesting that PINs selected for Signal are not widely shared with others and are considered confidential.

PIN Composition A participant's understanding of the Signal PIN's functionality had a large effect on the composition of their PIN. We asked participants what was their primary PIN selection strategy in **Q20**: code frequencies summarized in Figure 6 (with full details in Table 8 in Appendix D).

Among enthusiasts, password managers (PM) were mentioned frequently (28; 26 %). For example P100 noted that their "password safe generated it." Some participants mentioned the name of their password manager explicitly, like KeePass or Bitwarden. Far fewer casual Signal users (6; 7 %) mentioned a PM. The most-frequent code among casuals was *memorable*: (30; 36 %), choosing a PIN easy to remember; among enthusiasts it was second-most frequent (23; 21 %). For example, P7 noted their PIN was "Complicated enough but can still be remembered." This result suggests that despite the prevalence of randomized password generation, most participants want to select a PIN they can remember and recall easily, rather than having to look it up in a PM.

Interestingly, while the study of Markert et al. found dates to be the most-popular strategy for selecting a PIN, only 3 of our participants mentioned dates (2 enthusiasts and 1 casual) [25]. In the study of Markert et al. with (n = 200), *memorable* was the second-most frequent code (37; 19 %).

We then asked participants why they select a PIN with the current "security level" (**Q22**). (Results summarized in Figure 7; full details in Table 9 in Appendix D.) Among both enthusiasts (25; 23 %) and casual Signal (20; 24 %) users, many mentioned security; P44, an enthusiast, said "I am fairly security conscious."

Casuals and enthusiasts roughly equally mentioned that they chose something that was simply *good enough*: (16; 15 %) and (15; 18 %) respectively. Slightly more casual users mentioned memorability: (12; 11 %) enthusiasts and (18; 22 %) casual users. A similar number of enthusiasts (11; 10 %) and casuals (6; 7 %) mentioned that they try to be consistent in their security choices around PINs (and authentication generally), for example "Because I always choose this security level" (P109).

Recall that while Signal refers to this secret as a PIN, it is not a traditional *personal identification number*, but rather has more of the properties of a password. We asked participants to provide metrics for how many numbers, characters, and special symbols they use in their Signal PIN (**Q24**). Participants were presented a slider for each class from 0 to 12. While it is of course possible that a participant might have more than 12 digits, as a practical matter more than this simply indicates the use of a PM, which we can see in our data. Results are shown in Table 2.

Enthusiasts on average chose PINs with an additional 1.3 digits, 3.0 letters, and 1.3 special characters, and length increased overall by 5.5 characters. Except for the number of special characters, we were able to observe significant differences between the enthusiasts and the casuals using a t-test with Bonferroni-correction (for 8 overlapping hypotheses).

When dividing the population by their use of PMs, the difference is even greater. (Note that more enthusiasts employed a PM.) PM users chose PINs with an additional 2.1 digits, 5.3 letters, and 3.1 special characters. Overall, they used PINs which are 10.5 characters longer on average. Using a t-test with Bonferroni-correction, we were able to observe significant differences for all those statistics.

Table 2: PIN composition across different user groups $n = 191$ participants who set a PIN and did not disable it. t-tests were performed between groups within categories; all p-values are displayed Bonferroni-corrected for 8 overlapping hypothesis tests.

		Length		Digits		Letters		Special Characters	
Classification	**Participants**	**Mean (SD)**	**t-test**	**Mean (SD)**	**t-test**	**Mean (SD)**	**t-test**	**Mean (SD)**	**t-test**
Enthusiast	106	12.7 (9.8)	$t = 4.65$	6.2 (3.3)	$t = 2.97$	4.4 (5.1)	$t = 4.57$	2.2 (4.1)	$t = 2.74$
Casual	85	7.2 (5.7)	$p < 0.001^{**}$	4.9 (2.4)	$p = 0.026^{*}$	1.4 (3.3)	$p < 0.001^{**}$	0.9 (2.4)	$p = 0.05$
PM User	62	17.3 (10.2)	$t = 9.42$	7.0 (3.7)	$t = 4.72$	6.7 (5.1)	$t = 8.79$	3.7 (4.7)	$t = 6.16$
non-PM User	129	6.8 (5.3)	$p < 0.001^{**}$	4.9 (2.3)	$p < 0.001^{**}$	1.3 (3.2)	$p < 0.001^{**}$	0.6 (2.2)	$p < 0.001^{**}$
Overall	191	10.3 (8.7)	–	5.6 (3.0)	–	3.1 (4.7)	–	1.6 (3.5)	–

RQ3 Results Summary Many participants reuse Signal PINs in a number of ways. Roughly 15 % indicated that they use their Signal PIN as their screen lock PIN, used to unlock their smartphone. Nearly 30 % noted that the same PIN is used in other contexts, most commonly as an ATM/banking/payment card PIN. The Signal PIN is also reused in other mobile apps, such as a WhatsApp PIN, serving the same purpose as a Signal PIN for SVR and registration lock. When selecting a PIN, understanding of the purpose of Signal PINs led to much more diverse PINs, both in terms of the PIN length but also the presence of special characters and symbols. Among enthusiasts, the use of a password manager was particularly prominent when selecting a PIN, as compared to more casual users. But by far the largest factor in PIN selection overall is a desire for choosing a memorable PIN.

6 Discussion

Communicating about Signal PINs Our data show Signal's communication about the PIN feature has been effective for its traditional community of privacy enthusiasts. Without prompting, participants told us they learned about the PIN by reading blog posts, the Signal website, and tweets. Casual users, on the other hand, were much less likely to have exposure to these other sources. For this reason, in-app or in-the-moment resources nudging casual users in a more secure direction would almost certainly be of benefit.

As explained in Section 2, the case of Signal is especially challenging. While users are surely familiar with PINs as used in smartphone-unlock and payment-card scenarios, Signal PINs are actually used to *infrequently* derive encryption keys for SVR and *infrequently* act as a password for registration lock. Yet, despite the text in the Signal PIN enrollment prompt (see Figure 9) saying "You won't need your PIN to open the app," many of the participants who did not set a PIN mentioned inconvenience as a reason for their decision.

When further exploring the cause for this, the name "PIN" itself, is likely causing confusion. The Signal PIN is fundamentally a countermeasure against account takeover and to offer recovery functionality. If for example, the Signal PIN were to be called the "Account Recovery Password," or perhaps "Restore/Recovery Password," that might better convey the usage pattern. Text could then inform the user of the ill consequences of a bad PIN choice. This end could be achieved with text like "This password protects you from account takeover." Re-framing the PIN in this way could break the users' mental association with device-unlock PINs while also inspiring dread of consequences. While our study does not directly measure the effectiveness of such an intervention, the themes we uncovered naturally point in this direction.

Encouraging Password Managers The Signal PIN ultimately is used to derive a symmetric key in SVR and to retrieve a copy of the encrypted profile backup. For this reason alone, it is worth encouraging users to generate and store their Signal PIN in a password manager (PM). Few users are willing to memorize long, random keys and a PM is much better at generation, storage, and recall of secrets. Importantly, the user interface of a PM is already designed to explain these concepts to a user. The longest and most diverse PINs observed in the data were selected by participants using a PM.

But to reach this goal, broader adoption of PMs is also needed: while half of the enthusiasts in our study are already using a PM to manage their Signal PIN, only 10 casual participants do (10 %). For at least this group of users, this approach is preferable. The Signal app could reinforce this idea in the UI and encourage users to adopt a PM if they have not yet — and if they have, to use it to manage their Signal PIN.

PIN Security An account with a strong PIN is less likely to be taken over by an attacker on the network. Our data show large differences in how subgroups of participants select PINs. Although we did not ask participants for their Signal PIN, we asked for its composition among classes of characters: digits, letters, and special characters. Importantly Signal PIN security affects all users because account takeover can affect both the sender and receivers, especially in a group conversation. Even if a given user picks a strong PIN, if one of their messaging partners does not — that well-behaved user is at risk of mistakenly communicating sensitive data to an attacker who hijacked another account.

The current mechanisms of ensuring users select a strong PIN are minimal. Signal currently implements a very small blocklist of weak numeric PINs. These include the following: (a) not empty; (b) not sequential digits (e.g., 1234); (c) not all

the same digit (e.g., 0000) Note that this leaves other popular choices like recent years and dates as acceptable Signal PINs, which are often chosen by users [6, 25]. A targeted attack on an account where the victim's birthdate, anniversary, etc. are known would likely greatly assist the attacker. The sequence check also only applies to numeric PINs — observe "abcd" and "aaaa" are both valid PINs. In addition, this approach fails to block popular passwords like "password."

This situation could certainly be improved quite easily, for example implementing the blocklist as recommended by Markert et. al [25] and Bonneau et. al [6] for PINs and following recent guidance from the literature and from government agencies for passwords. NIST Special Publication 800-63B, recommends checking user password choices against lists of the most popular passwords [15]. PIN checks could easily occur locally on the user's device; however full password checks would require additional features to protect the privacy of the user's password.

PIN Verification Reminders To our knowledge, this is the highest-profile roll-out to date of PIN verification reminders (both on Signal and other messengers using the Signal protocol, like WhatsApp). While our study is based on user self-reported data, Figure 4 shows that participants do not generally feel they have a problem recalling their Signal PIN. This could be due to password manager use or that participants are using PINs they know well and use in other contexts. More than half of users say they frequently/very frequently verify their PIN when prompted, which points to user acceptance of PIN reminders. Even though (45; 24 %) of respondents turned off PIN reminders, many of those used a password manager; the remainder appear to be comfortable and appreciate periodic PIN verification.

7 Conclusion

We conducted an online study (n = 235) of Signal users recruited from Reddit, Signal Community Forum, snowballing, and Prolific about their understanding and choice of Signal PINs. In total, 86 % of participants set a PIN, with 57 % able to technically describe what Signal PINs are used for (enthusiasts) and 43 % unable to accurately describe how Signal PINs are used (casuals). We also find that PIN composition followed similar lines: enthusiasts use significantly longer PINs with more complex compositions, and casual participants used more traditional, numeric PINs despite the fact that Signal allows PINs to be alphanumeric. This suggests that communication about the Signal PIN has been effective for part of the Signal population only and that new strategies will be needed to reach the remainder.

As an example of in-app authentication — an authentication mechanism that occurs within a mobile app setting — our investigation shows that in the case of Signal, in-app usage of PINs can be confusing for users who have grown accustomed to screen lock and website login. These authentication metaphors are used often enough that users can be reasonably expected to handle them without much explanation. Where some authentication machinery (a PIN, for example) is repurposed for symmetric-key derivation, only enthusiasts can be expected to read the blogs, documents, tweets, and online help text to gain a full understanding.

Thus, we conclude that communication needs to meet the understanding of the (possibly multiple) user communities. Outside of a core constituency, even something as simple as the name matters. Signal's choice of the term "PIN" can be seen as correct and well-understood by the developers and enthusiasts. However, Signal may be well served in renaming their PIN, e.g., to "Account Recovery Password," and other uses of in-app authentication will need to carefully choose names and messaging to match user expectations.

Though our study does not measure the effect of this intervention, we believe there is strong evidence that suggests renaming Signal PIN to better reflect its usage could be helpful. First, a number of participants described it as an authentication mechanism or message privacy mechanism or simply indicated they do not know. A more precise name, like "Account Recovery," would help users place the Signal PIN in context with other credentials they manage. Second, reusing the term "PIN" suggests to users that only digits are valid. Using the word "Password" or "Passcode" could elicit broader classes beyond digits and encourage more diverse composition.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 184530. Further support was received through the research training group "Human Centered Systems Security" sponsored by the state of North Rhine-Westphalia, Germany, and the German Research Foundation (DFG) within the framework of the Excellence Strategy of the Federal Government and the States – EXC 2092 CASA – 390781972.

References

[1] Ruba Abu-Salma, M. Angela Sasse, Joseph Bonneau, Anastasia Danilova, Alena Naiakshina, and Matthew Smith. Obstacles to the Adoption of Secure Communication Tools. In *IEEE Symposium on Security and Privacy*, SP '17, pages 137–153, San Jose, California, USA, 2017.

[2] Nathanael Andrews. Can I Get Your Digits: Illegal Acquisition of Wireless Phone Numbers for SIM-Swap Attacks and Wireless Provider Liability. *Northwestern Journal of Technology and Intellectual Property*, 16(2):79–106, 2018.

[3] Adam J. Aviv, Katherine Gibson, Evan Mossop, Matt Blaze, and Jonathan M. Smith. Smudge Attacks on Smartphone Touch Screens. In *USENIX Workshop on Offensive Technologies*, WOOT '10, pages 1–7, Washington, District of Columbia, USA, 2010.

[4] Jeremiah Blocki, Saranga Komanduri, Lorrie Faith Cranor, and Anupam Datta. Spaced Repetition and Mnemonics Enable Recall of Multiple Strong Passwords. In *Symposium on Network and Distributed System Security*, NDSS '15, San Diego, California, USA, 2015.

[5] Joseph Bonneau. The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords. In *IEEE Symposium on Security and Privacy*, SP '12, pages 538–552, San Jose, California, USA, 2012.

[6] Joseph Bonneau, Sören Preibusch, and Ross Anderson. A Birthday Present Every Eleven Wallets? The Security of Customer-Chosen Banking PINs. In *Financial Cryptography and Data Security*, FC '12, pages 25–40, Kralendijk, Bonaire, 2012.

[7] Joseph Bonneau and Stuart Schechter. Towards Reliable Storage of 56-bit Secrets in Human Memory. In *USENIX Security Symposium*, SSYM '14, pages 607–623, San Diego, California, USA, 2014.

[8] Maria Casimiro, Joe Segel, Lewei Li, Yigeng Wang, and Lorrie Faith Cranor. A Quest for Inspiration: How Users Create and Reuse PINs. In *Who Are You?! Adventures in Authentication Workshop*, WAY '20, pages 1–7, Virtual Conference, 2020.

[9] Katriel Cohn-Gordon, Cas Cremers, Benjamin Dowling, Luke Garratt, and Douglas Stebila. A Formal Security Analysis of the Signal Messaging Protocol. *Journal of Cryptology*, 33(4):1914–1983, 2020.

[10] Sauvik Das, Tiffany Hyun-Jin Kim, Laura A. Dabbish, and Jason I. Hong. The Effect of Social Influence on Security Sensitivity. In *Symposium on Usable Privacy and Security*, SOUPS '14, pages 143–157, Menlo Park, California, USA, 2014.

[11] Sauvik Das, Adam D.I. Kramer, Laura A. Dabbish, and Jason I. Hong. Increasing Security Sensitivity With Social Proof: A Large-Scale Experimental Confirmation. In *ACM Conference on Computer and Communications Security*, CCS '14, pages 739–749, Scottsdale, Arizona, USA, 2014.

[12] Sauvik Das, Adam D.I. Kramer, Laura A. Dabbish, and Jason I. Hong. The Role of Social Influence in Security Feature Adoption. In *ACM Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW '17, page 1416–1426, Vancouver, British Columbia, Canada, 2017.

[13] Alexander De Luca, Sauvik Das, Martin Ortlieb, Iulia Ion, and Ben Laurie. Expert and Non-Expert Attitudes Towards (Secure) Instant Messaging. In *Symposium on Usable Privacy and Security*, SOUPS '16, pages 147–157, Denver, Colorado, USA, 2016.

[14] Robert J. Fisher. Social Desirability Bias and the Validity of Indirect Questioning. *Journal of Consumer Research,*, 20(2):303–315, 1993.

[15] Paul A. Grassi, James L. Fenton, and William E. Burr. Digital Identity Guidelines – Authentication and Lifecycle Management: NIST Special Publication 800-63B, 2017.

[16] Marian Harbach, Emanuel von Zezschwitz, Andreas Fichtner, Alexander De Luca, and Matthew Smith. It's a Hard Lock Life: A Field Study of Smartphone (Un)Locking Behavior and Risk Perception. In *Symposium on Usable Privacy and Security*, SOUPS '14, pages 213–230, Menlo Park, California, USA, 2014.

[17] Ann-Marie Horcher and Gurvirender P. Tejay. Building A Better Password: The Role of Cognitive Load in Information Security Training. In *IEEE International Conference on Intelligence and Security Informatics*, ISI '09, pages 113–118, Richardson, Texas, USA, June 2009.

[18] Roger Piqueras Jover. Security Analysis of SMS as a Second Factor of Authentication: The Challenges of Multifactor Authentication Based on SMS, Including Cellular Security Deficiencies, SS7 Exploits, and SIM Swapping. *ACM Queue*, 18(4):37–60, 2020.

[19] Hassan Khan, Jason Ceci, Jonah Stegman, Adam J. Aviv, Rozita Dara, and Ravi Kuber. Widely Reused and Shared, Infrequently Updated, and Sometimes Inherited: A Holistic View of PIN Authentication in Digital Lives and Beyond. In *Annual Computer Security Applications Conference*, ACSAC '20, pages 249–262, Austin, Texas, USA, 2020.

[20] Hyoungshick Kim and Jun Ho Huh. PIN Selection Policies: Are They Really Effective? *Computers & Security*, 31(4):484–496, 2012.

[21] Thomas K. Landauer and Robert A. Bjork. Optimum Rehearsal Patterns and Name Learning. In *International Conference on Practical Aspects of Memory*, PAM '78, pages 625–632, Cardiff, United Kingdom, 1978.

[22] Kevin Lee, Benjamin Kaiser, Jonathan Mayer, and Arvind Narayanan. An Empirical Study of Wireless Carrier Authentication for SIM Swaps. In *Symposium on Usable Privacy and Security*, SOUPS '20, pages 61–79, Virtual Conference, 2020.

[23] Joshua Lund. Technology Preview for Secure Value Recovery, 2019. https://signal.org/blog/secure-value-recovery, as of June 8, 2021.

[24] Eleanor E. Maccoby and Nathan Maccoby. *The Interview: A Tool of Social Science*, volume 1, pages 449–487. New York, USA, 1954.

[25] Philipp Markert, Daniel V. Bailey, Maximilian Golla, Markus Dürmuth, and Adam J. Aviv. This PIN Can Be Easily Guessed: Analyzing the Security of Smartphone Unlock PINs. In *IEEE Symposium on Security and Privacy*, SP '20, pages 1525–1542, San Francisco, California, USA, 2020.

[26] Moxie Marlinspike. Facebook Messenger Deploys Signal Protocol for End-to-End Encryption, 2016. https://signal.org/blog/facebook-messenger, as of June 8, 2021.

[27] Moxie Marlinspike. WhatsApp's Signal Protocol Integration Is Now Complete, 2016. https://signal.org/blog/whatsapp-complete, as of June 8, 2021.

[28] William Melicher, Darya Kurilova, Sean M. Segreti, Pranshu Kalvani, Richard Shay, Blase Ur, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Michelle L. Mazurek. Usability and Security of Text Passwords on Mobile Devices. In *ACM Conference on Human Factors in Computing Systems*, CHI '16, pages 527–539, San Jose, California, USA, May 2016.

[29] Arthur W. Melton. The Situation With Respect to the Spacing of Repetitions and Memory. *Journal of Verbal Learning and Verbal Behavior*, 9(5):596–606, 1970.

[30] Steven Mujye and Yair Levy. Complex Passwords: How Far Is Too Far? the Role of Cognitive Load on Employee Productivity. *Online Journal of Applied Knowledge Management*, 1(1):122–132, 2013.

[31] Christopher Novak, Jim Blythe, Ross Koppel, Vijay Kothari, and Sean Smith. Modeling Aggregate Security With User Agents That Employ Password Memorization Techniques. In *Who Are You?! Adventures in Authentication Workshop*, WAY '17, Santa Clara, California, USA, 2017.

[32] Sean Oesch, Ruba Abu-Salma, Oumar Diallo, Juliane Krämer, James Simmons, Justin Wu, and Scott Ruoti. Understanding User Perceptions of Security and Privacy for Group Chat: A Survey of Users in the US and UK. In *Annual Conference on Computer Security Applications*, ACSAC '20, pages 234–248, Virtual Conference, 2020.

[33] Jim Oleary. Improving Registration Lock with Secure Value Recovery, 2020. https://signal.org/blog/improving-registration-lock, as of June 8, 2021.

[34] Sarah Pearman, Shikun Aerin Zhang, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. Why People (Don't) Use Password Managers Effectively. In *Symposium on Usable Privacy and Security*, SOUPS '19, pages 319–338, Santa Clara, California, USA, 2019.

[35] Paul Pimsleur. A Memory Schedule. *The Modern Language Journal*, 51:73–75, 1967.

[36] Prolific Team. Deciding on a Reward, 2018. https://researcher-help.prolific.co/hc/en-gb/articles/360009500733-Deciding-on-a-Reward, as of June 8, 2021.

[37] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. How well do my results generalize? Comparing security and privacy survey results from MTurk, Web, and telephone samples. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 1326–1343. IEEE, 2019.

[38] Paul Rösler, Christian Mainka, and Jörg Schwenk. More is Less: On the End-to-End Security of Group Chats in Signal, WhatsApp, and Threema. In *European Symposium on Security and Privacy*, EuroSP '18, pages 415–429, London, United Kingdom, 2018.

[39] Scott Ruoti, Nathan Kim, Ben Burgon, Timothy van der Horst, and Kent Seamons. Confused Johnny: When Automatic Encryption Leads to Confusion and Mistakess. In *Symposium on Usable Privacy and Security*, SOUPS '13, pages 5:1–5:12, Newcastle, United Kingdom, 2013.

[40] Florian Schaub, Ruben Deyhle, and Michael Weber. Password Entry Usability and Shoulder Surfing Susceptibility on Different Smartphone Platforms. In *International Conference on Mobile and Ubiquitous Multimedia*, MUM '12, pages 13:1–13:10, Ulm, Germany, 2012.

[41] Stuart Schechter and Joseph Bonneau. Learning Assigned Secrets for Unlocking Mobile Devices. In *Symposium on Usable Privacy and Security*, SOUPS '15, pages 277–295, Ottawa, Ontario, Canada, July 2015. USENIX.

[42] Sunyoung Seiler-Hwang, Patricia Arias-Cabarcos, Andrés Marín, Florina Almenares, Daniel Díaz-Sánchez, and Christian Becker. "I Don't See Why I Would Ever Want to Use It" Analyzing the Usability of Popular Smartphone Password Managers. In *ACM Conference on Computer and Communications Security*, CCS '19, pages 1937–1953, London, United Kingdom, 2019.

[43] Manish Singh. Signal's Brian Acton Talks About Exploding Growth, Monetization, and WhatsApp Data Sharing Outrage, 2021. https://tcrn.ch/38BHusb, as of June 8, 2021.

[44] Herbert F. Spitzer. Studies in Retention. *Journal of Educational Psychology*, 30(9):641–656, 1939.

[45] U.S. Department of Homeland Security. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research, 2012. https://www.caida.org/publications/papers/2012/menlo_report_actual_formatted, as of June 8, 2021.

[46] Elham Vaziripour, Devon Howard, Jake Tyler, Mark O'Neill, Justin Wu, Kent Seamons, and Daniel Zappala. I Don't Even Have to Bother Them! Using Social Media to Automate the Authentication Ceremony in Secure Messaging. In *ACM Conference on Human Factors in Computing Systems*, CHI '19, pages 934:1–93:12, Glasgow, Scotland, United Kingdom, 2019.

[47] Elham Vaziripour, Justin Wu, Mark O'Neill, Daniel Metro, Josh Cockrell, Timothy Moffett, Jordan Whitehead, Nick Bonner, Kent Seamons, and Daniel Zappala. Action Needed! Helping Users Find and Complete the Authentication Ceremony in Signal. In *Symposium on Usable Privacy and Security*, SOUPS '18, pages 47–62, Baltimore, Maryland, USA, 2018.

[48] Elham Vaziripour, Justin Wu, Mark O'Neill, Jordan Whitehead, Scott Heidbrink, Kent Seamons, and Daniel Zappala. Is That You, Alice? A Usability Study of the Authentication Ceremony of Secure Messaging Applications. In *Symposium on Usable Privacy and Security*, SOUPS '17, pages 29–47, Santa Clara, California, USA, 2017.

[49] Ding Wang, Qianchen Gu, Xinyi Huang, and Ping Wang. Understanding Human-Chosen PINs: Characteristics, Distribution and Security. In *ACM Asia Conference on Computer and Communications Security*, ASIA CCS '17, pages 372–385, Abu Dhabi, United Arab Emirates, 2017.

[50] WhatsApp. Answering Your Questions About WhatsApp's Privacy Policy, 2021. https://faq.whatsapp.com/general/security-and-privacy/answering-your-questions-about-whatsapps-privacy-policy, as of June 8, 2021.

[51] Justin Wu, Cyrus Gattrell, Devon Howard, Jake Tyler, Elham Vaziripour, Daniel Zappala, and Kent Seamons. "Something Isn't Secure, but I'm Not Sure How That Translates Into a Problem": Promoting Autonomy by Designing for Understanding in Signal. In *Symposium on Usable Privacy and Security*, SOUPS '19, pages 137–156, Santa Clara, California, USA, 2019.

Appendix

A Additional Pre-Screening Study

The following question was asked in an additional pre-screening study on Prolific to be able to recruit more Signal users for our main study:

P1 Which instant messaging apps do you use? (Select all that apply)
□ WhatsApp □ Facebook Messenger □ Signal □ Telegram □ iMessage
□ WeChat □ QQ □ Other, please specify: _____

B Survey Instrument of the Main Study

Q1 Signal Private Messenger is a cross-platform encrypted messaging service. Do you use Signal?
∘ Yes ∘ No

[Participants who indicate No are screened out of the survey at this point, and only Signal users move forward]

Q2 I use Signal primarily on:
∘ Android ∘ Apple iPhone
∘ Other, please specify: _____

Q3 I also use Signal on: (Select all that apply)
□ Desktop □ Tablet □ None of these

Q4 PINs are a new feature provided by Signal. In your own words, please explain how PINs are used by Signal.
Answer: _____

Q5 Did you set a Signal PIN?
∘ Yes ∘ No

*[Participants who indicate Yes to **Q5**]*
Q6a Why did you choose to set a PIN?
Answer: _____

*[Participants who indicate No to **Q5**]*
Q6b Why did you choose not to set a PIN?
Answer: _____
*[Participants who indicate No to **Q5** skip ahead to **Q25**]*

Q7 Since setting your Signal PIN, are you still using it, or have you since disabled it?
∘ My Signal PIN is currently enabled ∘ My Signal PIN is currently disabled

*[Participants who indicated that their PIN is disabled in **Q7**]*
Q8 Why did you disable your Signal PIN?
Answer: _____
*[Participants who indicated that their PIN is disabled in **Q7** skip ahead to **Q25**]*

*[Participants who indicated that their PIN is enabled in **Q7**]*
Q9 How frequently do you have difficulty remembering your Signal PIN?
∘ Very frequently ∘ Frequently ∘ Occasionally ∘ Rarely ∘ Very rarely ∘ Never

*[Participants who indicated that their PIN is enabled in **Q7**]*
Q10 If you were to forget your Signal PIN, what would you do?
Answer: _____

[A screenshot of the Verify PIN prompt (see Figure 1)]

Q11 Have you seen this dialog in Signal?
∘ Yes ∘ No

*[Participants who indicated that they have seen the dialog in **Q11**]*
Q12 When prompted, how frequently do you verify your Signal PIN?
∘ Very frequently ∘ Frequently ∘ Occasionally ∘ Rarely ∘ Very Rarely ∘ Never

*[Participants who indicated that they have seen the dialog in **Q11**]*
Q13 Have you disabled Signal PIN reminders?
∘ Yes ∘ No

*[Participants who indicated that they have seen the dialog in **Q11** and that they have disabled reminders in **Q13**:]*

Q14 Why did you disable Signal PIN reminders?
Answer: _____

Q15 Many smartphone users also unlock their phone using a PIN or passcode. Is your Signal PIN the same one you use to unlock your smartphone?
∘ Yes ∘ No ∘ Unsure ∘ I do not lock my smartphone with a PIN or passcode

Q16 Do you use your Signal PIN in other contexts besides unlocking your smartphone? (Select all that apply)
□ ATM/Credit/Payment Card □ Laptop/PC □ Online Accounts
□ Electronic Door Lock □ Home Security System/Safe □ Garage Door Opener
□ Car/Truck/SUV □ Bike/Gym lock □ Voicemail □ Gaming Console
□ Smartwatch □ Other, please specify: _____

Q17 Do you use your Signal PIN in any other mobile applications?
∘ Yes, please specify: _____ ∘ No

Q18 Do you share your Signal PIN with friends or family?
∘ Yes ∘ No

Q19 How long is your Signal PIN?
Answer: _____

Q20 What was your primary strategy in selecting your Signal PIN?
Answer: _____

Q21 Compared to other PINs you use, did you try to pick a Signal PIN that was:
∘ The most secure PIN you use ∘ About the same security as other PINs you use ∘ Less secure than other PINs you use

Q22 Why did you choose a PIN with this security level?
Answer: _____

Q23 What is the shape of a red ball?
∘ Red ∘ Round ∘ Blue ∘ Square

[For each category, this question uses sliders so the user can choose a value between 0 and 12, or check the category's box for "Not applicable:"]

Q24 My Signal PIN contains:
Digits: _____
Letters: _____
Special characters: _____

Q25 Do you use other messenger services like: (Select all that apply)
□ Facebook messenger □ Skype □ Telegram □ WeChat □ WhatsApp
□ Other, please specify: _____
[For the services above, place them in order of how often you use them:]

Q26 Besides Signal, did you set a PIN in one or more other messengers?
∘ Yes ∘ No

*[Participants who indicate Yes to **Q26**]*

Q27a Why did you set a PIN in the other messenger(s)?
Answer: _____

*[Participants who indicate No to **Q26**]*

Q27b Why didn't you set a PIN in the other messenger(s)?
Answer: _____
*[Participants who indicate No to **Q26** skip ahead to **D1**]*

*[Participants who indicate Yes to **Q26**]*

Q28 In which other messenger(s) did you set a PIN? (Select all that apply)
□ Facebook Messenger □ Skype □ Telegram □ WeChat □ WhatsApp
□ Other, please specify: _____

*[Participants who indicate Yes to **Q26**]*

Q29 Did you re-use the same PIN with any of these other messengers?
∘ Yes ∘ No

*[Participants who indicate Yes to **Q29**]*

Q30a Why did you re-use the same PIN in another messenger?
Answer: _____

*[Participants who indicate No to **Q29**]*

Q30b Why didn't you re-use the same PIN in another messenger?
Answer: _____

D1 What is your age range?
∘ 18-24 ∘ 25-34 ∘ 35-44 ∘ 45-54 ∘ 55-64 ∘ 65-74 ∘ 75 or older ∘ Prefer not to say

D2 With what gender do you identify?
∘ Male ∘ Female ∘ Non-Binary ∘ Other ∘ Prefer not to say

D3 What is the highest degree or level of school you have completed?
∘ Some high school ∘ High school ∘ Some college ∘ Trade, technical, or vocational training ∘ Associate's Degree ∘ Bachelor's Degree ∘ Master's Degree ∘ Professional Degree ∘ Doctorate ∘ Prefer not to say

D4 What is your country of residence?
[Drop-down all countries]

D5 Does your educational background or job field involve IT?
∘ Yes ∘ No ∘ Prefer not to say

C Additional Figures

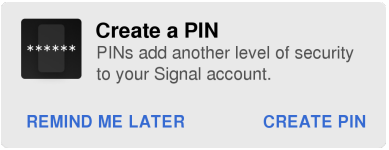


Figure 8: First prompt to ask Signal users to create a PIN.

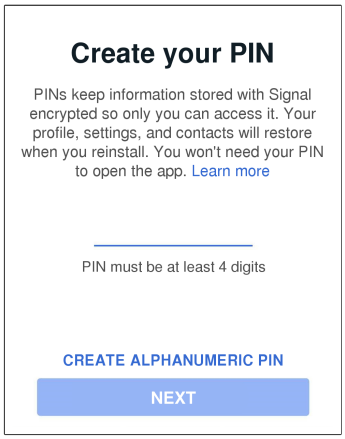


Figure 9: Updated prompt to ask Signal users to create a PIN.

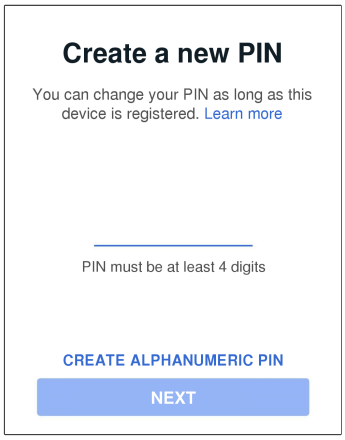


Figure 10: Prompt used when Signal users wish to change their PIN.

D Codebooks

We have 10 open-ended questions in our study for which two coders independently coded all answers we received. The two coders compared and combined codes until they agreed. For each question, *n* depicts the number of responses. As a single response might receive multiple codes, the number of codes does not sum to *n*. All codes of participant responses are shown below.

Table 3: **Q4**: "*PINs are a new feature provided by Signal. In your own words, please describe how PINs are used by Signal.*" (*n* = 235)

(a) Based on the answer to **Q4**, 132 participants were classified as *enthusiasts*.

Code Name	No.	%	Description	Sample from the Study
Backup	65	49 %	Participant mentions secure backup of settings and contacts but not messages	*"The PIN enables storing a backup of the user's settings on the signal servers in an encrypted form."* (P13)
Encryption	45	34 %	Participant mentions encryption based on the PIN	*"deriving a key to encrypt data stored on signals servers"* (P82)
Contacts	31	24 %	Participant mentions the backup of contact data	*"To secure contacts data saved on signal server with your own pin"* (P7)
Registration	23	17 %	Participant mentions the registration lock	*"to prevent reregistration of an account for the same mobile phone number for a given amount of time"* (P91)
Settings	8	6 %	Participant mentions the backup of settings	*"For encrypted backups - on cloud storage - for the user settings and profile. Not the messages themselves."* (P127)
Keying	7	5 %	Participant mentions the keying of the PIN	*"They say it's part of a keying mechanism providing a non-phone-number value that allows secure storage and retrieval of contacts and social graph info across devices."* (P2)
Phone number	6	5 %	Participant mentions the intention of Signal to move away from the phone number as an identifier	*"I think for backup purposes and to later fade out the phone number as identifier."* (P106)
Profile	4	3 %	Participant mentions the backup of profile information	*"PINs are used for recovery of settings and profile information after re-installation of Signal app."* (P54)
Groups	3	2 %	Participant mentions the backup of group memberships	*"They are used to secure private information such as group membership and store it on the Signal server'* (P35)
Anti-Cloud	2	2 %	Participant expresses negative sentiment about the data being stored by Signal	*"they are used to secure data in acloud service that is beeing forced on users"* (P208)
SVR	1	1 %	Participant mentions Secure Value Recovery (SVR)	*"Secure Value Recovery"* (P222)

(b) Based on the answer to **Q4**, 103 participants were classified as *casuals*.

Code Name	No.	%	Description	Sample from the Study
Don't Know	57	55 %	Participant does not mention any terms that may indicate an understanding	*"I don't understand their purpose very well. I thought that they might be using the PIN system to verify the identity of the person using signal (if for instance someone unauthorized gained access to the phone), but the way that pin entry is optionally offered every few weeks doesn't align with such a purpose. as such, I have no idea what they're trying to accomplish."* (P178)
Messages	21	20 %	Participant mentions the backup of messages	*"Secure backup of messages"* (P23)
Unlock	21	20 %	Participant mentions that the PIN is used to protect access to the app	*"Protect application from opening from an unlocked phone"* (P37)
Security	2	2 %	Participant mentions security	*"Security somehow...'* (P7)
Inconvenient	1	1 %	Participant mentions inconvenience	*"I have not tried it considering that it'd pop up for additional verification through the pin."* (P212)

Table 4: Codes assigned to the answers of the participants for (**Q6a**) and (**Q6b**) on adopting a PIN.

(a) **Q6a**: "*Why did you choose to set a PIN?*" (n = 202)

	Enthusiasts		Casuals			
Code Name	**No.**	**%**	**No.**	**%**	**Description**	**Sample from the Study**
Security	7	5 %	26	25 %	Participant mentions security	*"i wanted some extra security"* (P50)
Required	25	19 %	6	6 %	Participant mentions that there was no other choice	*"I did not see an option to not set one"* (P121)
Prompted	8	6 %	13	13 %	Participant mentions that Signal showed a prompt that suggested it	*"cause signal asked me to do so"* (P78)
Don't Know	4	3 %	16	16 %	Participant does not mention any of the terms that indicate an understanding	*"So that people that get a hold of my phone would have greater difficulty accessing my messages."* (P159)
Annoying	12	9 %	6	6 %	Participant mentions the feature was annoying	*"Because it kept hassling you with a pop up screen"* (P154)
Registration	14	11 %	2	2 %	Participant mentions the registration lock	*"I chose to set a PIN to both set registration lock and to backup my contacts."* (P51)
Features	8	6 %	3	3 %	Participant mentions features without further defining them	*"To be able to use the features that depend on a PIN"* (P111)
No harm	8	6 %	3	3 %	Participant describes there being no drawbacks	*"No disadvantage doing so"* (P20)
Trust	4	3 %	2	2 %	Participant expresses trust in Signal	*"I trusted the app and just did it when prompted."* (P155)
Privacy	2	2 %	3	3 %	Participant mentions valuing privacy	*"Because privacy is important to me and it's an added layer of it"* (P162)
Contacts	3	3 %	0	0 %	Participant mentions the backup of contact data	*"I want to be able to access contact data saved on signal server if I somehow can't access my current phone"* (P7)
Comfort	2	2 %	0	0 %	Participant mentions feeling comfortable	*"Because it I felt comfortable with the trade-off. Picked a long passphrase rather than a four digit PIN."* (P127)
Encryption	1	1 %	1	1 %	Participant mentions encryption based on the PIN	*"For me it's okay to encrypt and store data on Signal's servers as I have no high threat model."* (P87)
Lock	1	1 %	0	0 %	Participant mentions locking apart from registration lock	*"basically to lock and to avoid sim hijacking"* (P19)

(b) **Q6b**: "*Why did you choose not to set a PIN?*" (n = 33)

	Enthusiasts		Casuals			
Code Name	**No.**	**%**	**No.**	**%**	**Description**	**Sample from the Study**
Inconvenient	4	25 %	3	18 %	Participant mentions inconvenience	*"I want to access my apps as seamless and fast as possible."* (P212)
Anti-Cloud	7	44 %	0	0 %	Participant expresses negative sentiment about the data being stored by Signal	*"had no desire to have any contact data uploaded"* (P216)
Key management	3	19 %	0	0 %	Participant described the use of the PIN in key derivation	*"I don't trust Signal's encryption strategy involving SGX. It's my belief that SGX is likely to be compromised by nation-state actors, and cannot be used securely. If any of my private information must be stored persistently in a cloud service, it is unacceptable to use anything other than an encryption key that I personally control."* (P203)
Lock	0	0 %	4	24 %	Participant falsely links the phone lock to the PIN	*"My phone is always locked. Additional authentication seems unnecessary"* (P227)
No need	0	0 %	3	18 %	Participant mentions seeing no need	*"it's not necessary for me"* (P233)
Memorability	1	6 %	1	6 %	Participant described memorability issues	*"I didn't want to be bothered with remembering another code."* (P224)
No awareness	1	6 %	0	0 %	Participant did not know Signal had a PIN	*"I didn't know it existed."* (P232)
Not prompted	0	0 %	1	6 %	Participant said they were not prompted to set a PIN	*"was not asked."* (P218)
Rarely use	0	0 %	1	6 %	Participant described using Signal only rarely	*"I dont use signal much, its not for sensitive messages so dont need the extra security"* (P223)
Unsupported	0	0 %	1	6 %	Participant described using an unsupported client	*"Not possible because of using a unsupported native client for SailfishOS"* (P220)

Table 5: **Q8**: "*Why did you disable your Signal PIN?*" (n = 11)

	Enthusiasts		Casuals			
Code Name	**No.**	**%**	**No.**	**%**	**Description**	**Sample from the Study**
Annoying	3	60 %	1	17 %	Participant mentions being annoyed	*"It was annoying."* (P188)
Anti-Cloud	2	40 %	1	17 %	Participant expresses negative sentiment about the data being stored by Signal	*"Don't want my data stored on their server"* (P193)
Inconvenient	1	20 %	1	17 %	Participant mentions inconvenience	*"Verification overhead"* (P212)
No backup	1	20 %	0	0 %	Participant describes not needing a backup	*"It's annoying to re-enter the PIN and I don't need backup for signal since there's no important conversation"* (P231)
No need	1	20 %	1	17 %	Participant sees no necessity	*"I do not need it"* (P206)

Table 6: **Q10**: "*If you were to forget your Signal PIN, what would you do?*" (n = 191)

Code Name	Enthusiasts No.	Enthusiasts %	Casuals No.	Casuals %	Description	Sample from the Study
Don't know	27	25 %	33	40 %	Participant does not know what to do	*"Honestly don't know"* (P68)
PW Manager	45	42 %	12	15 %	Participant has the PIN stored in a password manager	*"I've stored my Signal PIN in my PW manager"* (P74)
Reset	0	0 %	12	15 %	Participant describes resetting the account	*"Check the help page for how to reset"* (P158)
Wait	8	7 %	4	5 %	Participant is aware that the PIN expires and would wait	*"wait for pin expiration"* (P161)
New PIN	4	5 %	7	6 %	Participant would set a new PIN	*"as long as I have access to my Signal account I can set a new PIN at any time"* (P18)
New account	4	4 %	5	6 %	Participant would create a new account	*"I would make another account"* (P181)
Reused	2	2 %	4	5 %	Participant reuses the PIN and does not expect to forget it	*"It is a PIN I use for my bank cards, so I would not forget it."* (P145)
Unrecoverable	2	2 %	3	4 %	Participant accepts that there is not way to recover	*"Signal said there is no way to recover it. All chats constants block list will be lost."* (P79)
Contact	0	0 %	4	5 %	Participant would contact Signal directly	*"Contact the signal team"* (P137)
Guess	0	0 %	3	4 %	Participant would try to guess the PIN	*"try a lot of PINs i use"* (P98)
Reinstall	2	2 %	0	0 %	Participant would reinstall Signal	*"delete the app and reinstall it"* (P106)
Written	1	1 %	1	1 %	Participant mentions that the PIN has been written down	*"I would check the PIN on my journal, I wrote it down with all the passwords and the login infos."* (P143)

Table 7: **Q14**: "*Why did you disable Signal PIN reminders?*" (n = 45)

Code Name	Enthusiasts No.	Enthusiasts %	Casuals No.	Casuals %	Description	Sample from the Study
PW Manager	22	67 %	1	8 %	Participant has the PIN stored in a password manager	*"Because I have a password safe and do not need to remember the pIn"* (P49)
Annoyed	6	18 %	5	42 %	Participant describes being annoyed	*"Because it asked my pin to often"* (P70)
No need	5	15 %	4	33 %	Participant describes not needing them	*"I dont think I need them"* (P160)
Memorized	0	0 %	1	8 %	Participant does not expect to forget the PIN	*"Thought I'd be able to remember it"* (P157)
Effective	0	0 %	1	9 %	Participant mentions the effectiveness of the reminders	*"After a few reminders I was sure not to forget the PIN"* (P87)

Table 8: **Q20**: "*What was your primary strategy in selecting your Signal PIN?*" (n = 191)

Code Name	Enthusiasts No.	Enthusiasts %	Casuals No.	Casuals %	Description	Sample from the Study
Memorable	23	21 %	30	36 %	Participant mentions memorability	*"My ability to remember it."* (P116)
PW Manager	28	26 %	6	7 %	Participant describes using a password manager	*"My password safe generated it."* (P100)
Reuse	16	15 %	13	16 %	Participant describes reusing a PIN	*"I used my PIN that I often use."* (P176)
Random	15	14 %	7	8 %	Participant describes choosing a random PIN	*"random number generator"* (P63)
Meaning	6	6 %	6	7 %	Participant describes choosing a meaningful PIN	*"Something meaningful to me"* (P77)
Security	3	3 %	8	10 %	Participant describes selecting a secure PIN	*"just something safe an long"* (P200)
Pattern	3	3 %	4	5 %	Participant describes choosing a PIN that depicts a pattern	*"Thinking of a pattern thats memorable to me"* (P142)
None	2	2 %	3	4 %	Participant describes not having a strategy	*"no strategy"* (P115)
Word	2	2 %	3	4 %	Participant describes converting a word to a PIN (textonyms)	*"Words to numbers"* (P115)
Date	2	2 %	1	1 %	Participant describes using a date	*"It's a date that is relevant but nobody knows"* (P154)
System	2	2 %	1	1 %	Participant describes having a certain system	*"My prefered format"* (P138)
Typable	0	0 %	1	1 %	Participant mentions a PIN that is easy to enter	*"Strong alphanumeric password that is secure enough but fairly easy to type on the phone, even if I couldn't paste it from password manager for some reason."* (P54)
Simple	0	0 %	1	1 %	Participant mentions simplicity	*"Something simple"* (P154)
Phone	0	0 %	1	1 %	Participant mentions a phone number	*"Old phone number i can remembee"* (P108)

Table 9: **Q22**: "*Why did you choose a PIN with this security level?*" (n = 191)

Code Name	Enthusiasts No.	Enthusiasts %	Casuals No.	Casuals %	Description	Sample from the Study
Memorability	12	11 %	18	22 %	Participant mentions memorability	*"Because I wanted it to be easy to remember."* (P5)
Enough	16	15 %	15	18 %	Participant describes the security level being sufficient	*"I think that's enough"* (P179)
Security	25	23 %	20	24 %	Participant mentions security	*"I am fairly security conscious"* (P44)
Consistent	11	10 %	6	7 %	Participant describes this level being the standard	*"Because I always choose this security level."* (P109)
Trade-off	9	8 %	2	2 %	Participant describes some form of trade-off	*"trade-off between remembering and security"* (P60)
Reuse	7	7 %	2	2 %	Participant describes reusing a PIN	*"The same as the iPhone passcode."* (P144)
PW manager	6	6 %	2	2 %	Participant describes using a password manager	*"why not, if i can use a pw manager"* (P76)
Don't know	1	1 %	6	7 %	Participant cannot remember the strategy	*"I don't remember"* (P84)
None	2	2 %	4	5 %	Participant describes not having a strategy	*"no strategy"* (P88)
Convenience	3	3 %	1	1 %	Participant mentions convenience	*"Convience over security"* (P113)
Privacy	2	2 %	1	1 %	Participant mentions privacy	*"The chats and contacts in Signal have a relatively high level of privacy, so it should be properly protected. Yet the pin is not as good as for example my computers encryption password but as good as my android encryption phrase."* (P94)
Low-threat	0	0 %	2	2 %	Participant sees little need for data security	*"The info isn't super important"* (P168)
Indifference	2	2 %	0	0 %	Participant says the PIN is unimportant	*"Dont think that the pin is too important"* (P120)
Rarely use	2	2 %	0	0 %	Participant described using the PIN only rarely	*"Unlike my smartphone unlock pin for example, I don't have to enter my Signal PIN frequently (never really, unless I set up a new smartphone) and thus had no problem with selecting a long and complicated PIN"* (P20)
Minimum	1	1 %	0	0 %	Participant mentions a Signal requirement	*"Initially 6 digits were required."* (P132)

Table 10: Codes assigned to the answers of the participants for (**Q27a**) and (**Q27b**) on setting a PIN in other messengers.

(a) **Q27a**: "*Why did you set a PIN in other messenger(s)?*" (n = 49)

Code Name	Enthusiasts No.	Enthusiasts %	Casuals No.	Casuals %	Description	Sample from the Study
Security	24	67 %	8	62 %	Participant mentions security	*"For more security, 2FA"* (P95)
Prompted	4	11 %	1	8 %	Participant mentions being prompted by the application	*"Prompted to do so, and I understand the reasons why it is a good idea."* (P196)
Required	4	11 %	1	8 %	Participant mentions that there was no other choice	*'Forced to set"* (P93)
Feature	2	6 %	1	8 %	Participant mentions being given the option to	*"Because I could"* (P117)
Don't know	1	3 %	2	16 %	Participant doesn't address the question	*"Telegram"* (P48)
Reuse	1	3 %	0	0 %	Participant mentions reusing a PIN when possible	*"Since I already has a pin memorized, why not use it in other messengers"* (P50)

(b) **Q27b**: "*Why didn't you set a PIN in other messenger(s)?*" (n = 131)

Code Name	Enthusiasts No.	Enthusiasts %	Casuals No.	Casuals %	Description	Sample from the Study
No feature	20	35 %	24	33 %	Participant mentions not being able to set a PIN	*"They don't have that option"* (P35)
No need	18	32 %	12	17 %	Participant describe that there is no necessity	*"Not required"* (P156)
Not asked	10	17 %	20	28 %	Participant describes not being asked to	*"Was not asked to"* (P67)
Use rarely	3	5 %	4	6 %	Participant describes only using them rarely	*"I don't use them often, if at all."* (P51)
Screen lock	3	5 %	2	3 %	Participant describes that the phone lock is sufficient	*"The phone in itself has a pin"* (P194)
Annoyed	2	3 %	1	2 %	Participant describes being annoyed	*"They are inconvenient, do not know how, and I do not use them for secure messaging. My Signal is already password protected so a pin seems redundant."* (P55)
Insecure	0	0 %	2	3 %	Participant describes that they don't use them for secure communication	*"Not intended for secure communication."* (P62)
Comfort	1	2 %	0	0 %	Participant mentions feeling comfortable	*"comfort"* (P102)

Table 11: Codes assigned to the answers of the participants for (**Q30a**) and (**Q30b**) on reusing the Signal PIN in another messenger.

(a) **Q30a**: "*Why did you re-use the same PIN in another messenger?*" (n = 10)

	Enthusiasts		Casuals			
Code Name	**No.**	**%**	**No.**	**%**	**Description**	**Sample from the Study**
Memorability	5	63 %	2	100 %	Participant mentions memorability	*"I was too lazy to memorize a new one... not good I know"* (P50)
Messenger PIN	2	25 %	0	0 %	Participant mentions using a PIN for messengers	*"Because I have one pin for messengers."* (P5)
Convenience	1	12 %	0	0 %	Participant mentions convenience	*"Convivence, but it was probably a poor decision, as WhatsApp is more vulnerable to a secret warant."* (P196)

(b) **Q30b**: "*Why didn't you re-use the same PIN in another messenger?*" (n = 37)

	Enthusiasts		Casuals			
Code Name	**No.**	**%**	**No.**	**%**	**Description**	**Sample from the Study**
Security	17	65 %	6	60 %	Participant mentions security	*"Reusing PINs is a bad practice."* (P54)
PW Manager	8	31 %	2	20 %	Participant describes using a password manager	*"Why would i? Thats what passwordmanagers are for.d"* (P23)
Other options	3	12 %	0	0 %	Participant describes having other options	*"Some of them gave me the option of using my thumbprint."* (P3)
Don't know	1	4 %	1	10 %	Participant cannot explain the reason	*"I didn't really think about it, it just happened"* (P179)
Required	0	0 %	1	10 %	Participant mentions different requirements	*"different lengths"* (P381)

On the Limited Impact of Visualizing Encryption: Perceptions of E2E Messaging Security

Christian Stransky[†], Dominik Wermke[C], Johanna Schrader[†], Nicolas Huaman[C], Yasemin Acar[‡], Anna Lena Fehlhaber[†], Miranda Wei[*], Blase Ur[◇], Sascha Fahl[†C]
[†] *Leibniz University Hannover;* [C] *CISPA Helmholtz Center for Information Security;*
[‡] *Max Planck Institute for Security and Privacy;* [*] *University of Washington;* [◇] *University of Chicago*

Abstract

Communication tools with end-to-end (E2E) encryption help users maintain their privacy. Although messengers like WhatsApp and Signal bring E2E encryption to a broad audience, past work has documented misconceptions of their security and privacy properties. Through a series of five online studies with 683 total participants, we investigated whether making an app's E2E encryption more visible improves perceptions of trust, security, and privacy. We first investigated why participants use particular messaging tools, validating a prior finding that many users mistakenly think SMS and e-mail are more secure than E2E-encrypted messengers. We then studied the effect of making E2E encryption more visible in a messaging app. We compared six different text disclosures, three different icons, and three different animations of the encryption process. We found that simple text disclosures that messages are "encrypted" are sufficient. Surprisingly, the icons negatively impacted perceptions. While qualitative responses to the animations showed they successfully conveyed and emphasized "security" and "encryption," the animations did not significantly impact participants' quantitative perceptions of the overall trustworthiness, security, and privacy of E2E-encrypted messaging. We confirmed and unpacked this result through a validation study, finding that user perceptions depend more on preconceived expectations and an app's reputation than visualizations of security mechanisms.

1 Introduction

The use of E2E-encrypted communication tools for e-mail (e.g., PGP [70], S/MIME [52]) or for mobile apps (e.g., WhatsApp [66], iMessage [6], Signal [56]) is an effective countermeasure against cybercriminals, nation-state attackers, and other adversaries [36]. Most E2E-encrypted communication tools provide confidentiality, integrity, authenticity, and perfect forward secrecy [20] for message contents, but do not hide metadata like sender/receiver identities or when the message was sent [46]. Many previous studies have documented usability and adoption challenges for encryption tools [8, 12, 15, 62], especially for e-mail encryption [25, 26, 47, 49] and modern E2E-encrypted messaging apps [3, 4].

Of particular concern is that users often have flawed mental models of E2E-encrypted tools' security and privacy properties. This can lead users to mistakenly use less secure alternatives like SMS or e-mail for confidential conversations even when they already have access to E2E encryption through widely used tools like WhatsApp and iMessage [3].

Recent work has highlighted how increasing the visibility of typically invisible security mechanisms can improve user perceptions of trust and security. For example, a qualitative study on e-voting found that displaying security mechanisms improved both user experience and need fulfillment [18]. In the context of E2E encryption on Facebook, another study's qualitative results suggested that visibly transforming Facebook messages to and from ciphertext (an implementation artifact in that work) appeared to increase user trust and perceptions of security [21]. For e-mail security, studies found that clearly labeling PGP-encrypted e-mail differently from unencrypted e-mail improved usability and perceived security, as well as reduced unintentional human error when interacting with PGP-encrypted e-mails [48, 50]. Our work tests these promising results in the space of mobile messaging apps. In an attempt to improve user comprehension and perceptions of security, privacy, and trust for E2E-encrypted mobile messaging apps, we thus investigated visualizing encryption through various text descriptions, icons, and animations of the encryption process.

We conducted a series of five user studies on MTurk and Prolific to investigate the following three research questions:

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

RQ 1: Which messaging tools do people prefer for confidential communications, and why?

Of participants who had an E2E-encrypted tool installed (80), 62.50% reported they would use a tool without E2E encryption for confidential conversations, echoing prior work [3]. This finding suggests that E2E-encrypted communication tools can do more to discourage users from switching to less-secure tools in situations when security and privacy matter. Our root-cause analysis revealed factors like specific UI features, trust in companies, and security misconceptions contributed to participants' decisions.

RQ 2a: How does visualizing encryption through text, icons, or animations impact perceptions of E2E-encrypted messaging tools' security, trust, and privacy?

RQ 2b: What external factors and expectations mediate encryption visualizations' impact on user perceptions?

In an attempt to highlight tools' E2E encryption, we investigated three types of visualizations: *text disclosures*, *icons*, and *animations*. In our remaining online studies, we investigated the impact of different variants of these visualizations. While some of these disclosures have been investigated previously, the animations are especially novel, as is our application of a consistent human-subjects protocol to study all three types.

We found that perceptions of a tool's security, trust, and privacy increased as soon as there was a simple indicator of encryption, such as a text statement that messages are encrypted (similar to WhatsApp's current interface). Contradicting the recent literature, additional emphasis did not appear to have much impact. More concretely, heavyweight animations and icons did not appear to provide much benefit beyond a lightweight text disclosure in emphasizing E2E-encrypted messengers' security properties to users. While qualitative data suggested that rich visualizations like animations successfully emphasized security and encryption, they did not significantly impact quantitative measures of user perception. Notably, much of the recent literature relies on qualitative observations, whereas our dual use of both perspectives highlights limitations of visualizing security. Through a final study with additional questions, we validated the surprising lack of a quantitative effect and further unpacked users' expectations.

In this paper, we make the following contributions:

- We detail which E2E-encrypted communication tools participants use in different situations, and why.
- We investigate how visualizing encryption through text disclosures, icons, and animations impacts perceptions of security, privacy, and trust.
- We validate our findings and unpack the limitations of visualizing E2E encryption.

The rest of the paper is structured as follows. Section 2 presents previous work relevant to this paper and illustrates the novelty of our research. Section 3 provides detailed information on our methodology, including data quality and data analysis techniques we applied, as well as the ethical considerations and limitations of our work. In Section 4, we discuss the procedure and findings of our first study on the use of communication tools. Section 5 gives a detailed overview of the experiments we conducted on different visualizations of encryption, and Section 6 describes a validation study. Section 7 discusses our results, highlights their implications for secure messaging applications, and outlines possible future work. Finally, we conclude in Section 9.

2 Related Work

We discuss related work on encrypted communication tools' usability, adoption, and perception, and previous attempts to visualize security, especially encryption.

Usability of E2E Encryption The usability of E2E encryption has been a research focus since at least 1999, when Whitten and Tygar evaluated PGP with cognitive walkthroughs in a landmark paper [67]. One-third of participants failed to sign and encrypt an e-mail message within 90 minutes.

More recent work observes similar barriers. In two-person lab sessions, Ruoti et al. examined initial user experiences for three secure e-mail systems (Pwm, Tutanota, Virtru) through role-play scenarios with 50 participants. They found that participants were interested in secure e-mail in the abstract, but were unsure when and how they actually would use it. Only a few participants desired to use secure e-mail regularly [47]. De Luca et al. conducted online studies and interviews to investigate the role of security and privacy in people's decisions to use secure messaging apps. They reported that peer influence primarily drove decisions to use a particular secure messaging app; security and privacy were minor factors [14].

A number of prior research studies utilized interviews [4, 5, 9, 27, 68] or surveys [3, 5] to investigate users' mental models of E2E encryption. Similar to the findings of our first of five studies, these works identified a number of misconceptions regarding the security properties of E2E encryption. We based some of our survey questions on this prior work in an attempt to gain deeper insight into the root causes of users' security misconceptions and to try to mitigate such misconceptions.

Visualizing Encryption We discuss literature on visualizing encryption in three areas: web, e-mail, and messaging.

Visualizing and highlighting whether or not webpages are SSL/TLS-encrypted was historically a major focus of usable security research [58, 61]. In a lab setting, Whalen et al. conducted an eye-tracking study with 16 participants to test visual cues for SSL warnings, finding that icons provide prominent visual cues, yet they must be large and prominently placed [65]. Accordingly, we designed sufficiently large cues and placed them prominently in the center of our messaging app. Both Sobey et al. [57] and Maurer et al. [35] investigated alternative display methods, including full-browser themes, as security indicators of extended validity certificates. They

found that additional indicators of the level of security improved user confidence, the ease of finding information, and user understanding. Based on their work, we tested a number of variations for each type of visual cue. Schechter et al. conducted a qualitative lab study with 67 participants about the effect of removing security indicators on a banking website [54], finding that users ignore security indicators and that study designs incorporating role-playing reduce participants' security behaviours. More recently, in 2016 Felt et al. conducted a large quantitative online survey with 1329 participants, testing multiple cryptography-related labels and icons. They arrived at three indicators consisting of icons and labels for valid and invalid HTTPS and HTTP certificates to visualize the security level of the connection [23]. We built on this prior work by applying a similar but extended approach to the area of encrypted messaging apps, including the addition of qualitative elements and a validation study.

In the context of encrypting e-mail, related work investigates how user errors can be prevented and perceptions of security can be improved using security indicators. Two recent connected studies from 2013 and 2015 by Ruoti et al. proposed a web interface to support PGP encryption [48, 50]. They found that visualizing encryption using labels and adding scrambled text as an indicator of encrypted text helped to reduce user error when using PGP and supports trust in e-mail encryption. They proposed to further improve trust by letting users copy and paste e-mail ciphertext, but in a followup study found that doing so had no measurable effect on usability or security perceptions. Garfinkel et al. found that Key Continuity Management (KCM) systems with color-coded messages could improve e-mail security and effectively help novice users identify signed e-mails [26]. In 2015, Atwater et al. conducted a lab study investigating how a web interface can support e-mail encryption [8]. They found that participants prefer PGP to be integrated into their existing tool (e.g., Gmail). Participants' trust perceptions were based not on the tool's design, but rather the tool's overall reputation. Based on this finding that encryption should integrate into existing and well-known tools, we chose to test our own indicators using a modified version of the highly popular, E2E-encrypted WhatsApp Messenger.

Finally, we discuss related work regarding instant messaging and mobile apps. In 2012, Fahl et al. designed a tool for E2E encryption of private Facebook messages, evaluating the tool through lab and interview studies [21]. An artifact of their tool's implementation was that participants would see plaintext Facebook messages being translated to and from ciphertext. Their qualitative results implied that participants seeing the ciphertext upon sending or receiving messages was viewed positively and seemed to increase trust in the tool's security properties. In a lab study of the SELENE electronic voting protocol, Distler et al. [18] investigated how users reacted to seeing an explanation of encryption during the voting process. They found that overall perspicuity and users' perceptions of security increased due to the added waiting screen. In a followup online survey [17], they also tested different wordings of encryption in the scenarios of e-voting, online pharmacies, and online banking. They concluded that explanations of encryption should consist of short text without many elements, underpinning the design of the text disclosures we tested in Section 5.2.

In 2018, Demjaha et al. conducted an online study with 96 participants investigating metaphors to explain E2E encryption to users [16]. They concluded that wordings like "encryption" might be overloaded for end users and alternative metaphors might better explain the strengths and weaknesses of E2E encryption. While we focus on differences in structural explanations, we implement some metaphorical approaches in our icons and animations, measuring their effects compared to more straightforward labels and icons. Schröder et al. investigated authenticity-related error messages for the Signal [56] Android app [55]. They conducted a mostly qualitative study with 28 participants, finding that Signal needs to improve the awareness and verification of authenticity in conversations, as well as to communicate risks more clearly (e.g., providing guidelines for handling potential MITM attacks). Their findings suggest that the security perceptions of Signal could be improved in general.

In a recent study Akgul et al. evaluated if in-workflow messages in a messenger could improved the mental models of E2E encryption and found that while participants noticed them, they did not pay much attention to it, which limited the effect [5].

3 Methodology

We conducted a series of online studies on MTurk and Prolific (cf. Figure 1). This section gives a high-level overview of our approach. Section 4 details our study investigating current use of communication tools. Section 5 describes our studies on how different designs of encryption visualizations impact user perceptions.

Overall, we conducted five different user studies with 683 participants. For the first four, we recruited on MTurk. For the fifth, which was our validation study, we recruited on Prolific. We required participants in studies 2–5 be experienced WhatsApp users, enforcing this requirement through a qualification task on MTurk (cf. Section 5) and Prolific's built-in participant filters. We decided to use WhatsApp for our studies, since it is the most commonly used messenger with E2E encryption enabled by default in the US that is available on multiple platforms [59]. We estimated required participant numbers for each survey using power analysis and were limited by the total number of available WhatsApp users.

Each study had a distinct purpose:

Study 1: Use of Communication Tools The purpose of this study was to gain insight into the selection of communication tools for both day-to-day and confidential conversations. We aimed to understand how and why users decide to use certain

tools in particular circumstances. Table 4 illustrates messengers that were considered in this paper, and their features. Based on previous work [26, 50, 67], the results of Study 1, and the visual design of modern secure messaging apps, we then implemented potential encryption visualizations in a modern secure messaging app. Our goal was to investigate whether adding encryption visualizations to E2E-encrypted messaging app's UI would increase perceptions of trust, security, and privacy without sacrificing usability. The results for this study can be found in section 4.

Study 2: Disclosures Current secure messaging apps use specific textual framing (disclosures) to inform their users that conversations are E2E-encrypted. For example, WhatsApp displays *"Messages to this chat and calls are now secured with end-to-end encryption."*. However, prior studies on private browsing modes [69] and security warnings [22] have illustrated users' confusion about analogous disclosures. Therefore, we aimed to investigate whether a more detailed and technically correct ("end-to-end encrypted") disclosure had a different contribution to perceived security than more generous disclosures that are still connected to messaging security and comparatively tested six different versions. The results for this study can be found in section 5.2.

Study 3: Icons In addition to disclosures, a common approach is the use of security icons (e. g., lock symbols) [23] to indicate the presence of encryption or other security mechanisms. Similar to Study 2, we based our analysis on current secure messaging apps' security icons and icons discussed in previous usable security papers [23, 54]. We investigated three different icons, studying their impact on perceived trust, security and privacy, and usability. The results for this study can be found in section 5.3.

Study 4: Animations Additionally, we implemented and studied three animations of encryption. Prior work [18, 21] and the results of Study 1 implied that dynamic animations of the encryption process (e. g. disappearing messages or animations of plaintext turning into ciphertext) might increase perceptions of trust, security, and privacy. The results for this study can be found in section 5.4.

Study 5: Validation To validate and clarify the findings from studies 1–4, we performed a fifth study that addresses limitations of the previous four. One key challenge of studies 1–4 is the demographic bias of Amazon MTurk. Recent research identified generalizability and data quality issues on MTurk [31]. To account for this, we switched recruitment platforms, choosing Prolific [42]. Prolific provides strong tools to obtain a more diverse sample. Additionally, we performed the validation study to investigate root causes of particular results of Studies 2–4, so we also added an additional control condition and qualitative questions. To remove a potential confound suggested by the results of Studies 2–4, we also changed the messaging app from WhatsApp to a fictitious app we called Erebus. The results for this study can be found in section 6.

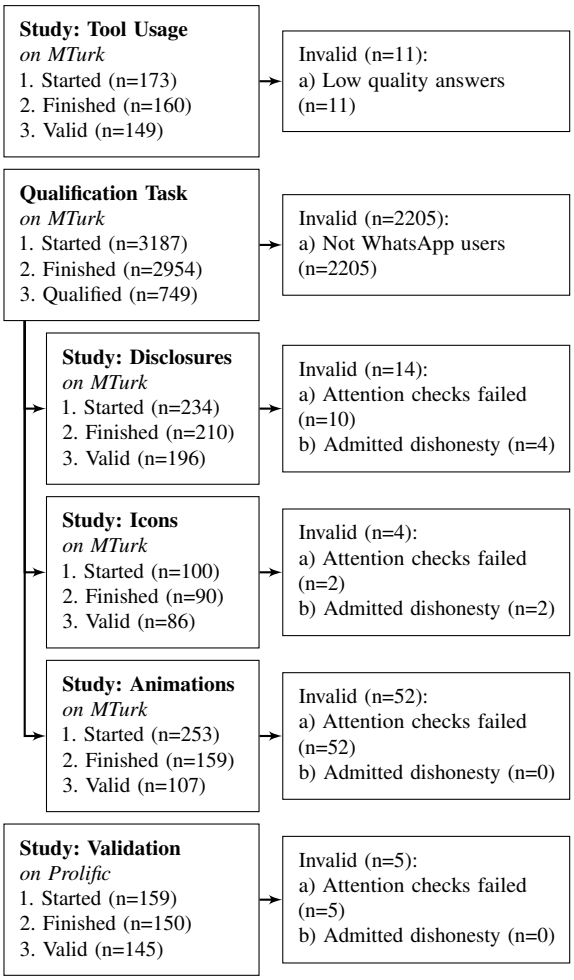


Figure 1: Illustration of our research procedure including *survey platform*, number of participants, and dropouts.

3.1 Study Procedure

We conducted all five studies sequentially to allow the findings of preceding studies to inform the design of later studies. For example, we used the most promising text disclosure from Study 2 in Studies 3–5.

Lab vs. Online Study Across studies 2–5 we investigated six different text disclosures (cf. Section 5.2), three different icons (cf. Section 5.3) and three different dynamic animations (cf. Section 5.4). Consequently, we recruited a rather high number of participants ($n = 534$ total participants). This made a laboratory experiment infeasible. Hence, we decided to conduct our experiments online using Amazon Mechanical Turk and Prolific Academic. Both platforms are popular amongst usable security and privacy user studies [1, 29, 37, 63].

Mockups vs. Real App We aimed for high internal validity to ensure that font sizes, types, positions of icons, animations, and the content of conversations (cf. Figure 7) remained consistent for all participants. Hence, we decided to use mockups

Factor	Description
Required	
Condition	Disclosures, icons, or animations (baseline: Control)
Optional	
CS Edu	Has CS education (self-reported, baseline: No)
CS Job	Has CS job (self-reported, baseline: No)
Age	Age in years (self-reported)

Table 1: Factors used in regression models. Model candidates were defined using all possible combinations of optional factors, with the required factors included in every candidate. Final models were selected by minimum AIC. Categorical factors are individually compared to the baseline.

instead of asking participants to install a real app on their devices. For the mockups, we created screencasts by forking the Signal Android app [56], since it implements the same encryption workflow that WhatsApp uses. We implemented the WhatsApp look and feel and all encryption visualizations. We recorded screencasts using the app and the conversation in Figure 7. During the conversation, we presented the different visualizations in each condition.

Additionally, each study had an online survey questionnaire at the end. The survey questionnaire addressed the perceived usability, trust, security, privacy and satisfaction with the tool, tool preference for both day-to-day and confidential conversations and demographic information about our participants.

Because we expected significant learning effect across conditions, Studies 2–5 followed a between-groups design.

Pre-Testing Before we conducted the studies, we pre-tested our questionnaires and screencasts, following best practices for cognitive interviews [24]. To glean insights into how survey respondents might interpret and answer questions and how they perceive the screencasts, we asked participants to share their thoughts as they answered each survey question and watched the screencasts. We used the findings to iteratively revise and rewrite our survey questions to minimize bias and maximize validity and modify the screencasts based on the feedback. We conducted cognitive interviews with members of our research group and university students, and performed a pre-test on MTurk to evaluate our survey questions under realistic conditions and to calibrate compensation relative to the time required. Pilots took an average of 15 minutes, so we compensated participants $2.50 (an hourly wage of $10).

Data Analysis Prior to data analysis, we took measures to ensure data quality (cf. A.5).

We perform both quantitative and qualitative data analysis. Throughout the paper, we measure usability using the UMUX Lite questionnaire [34]. We compare responses to the UMUX Lite across conditions with Pearson's chi-squared test (χ^2). We also collect net promoter scores, which are a quantitative measure of willingness to recommend a product. As these scores are continuous, we use the non-parametric Kruskal-Wallis H test (KW-H) for comparing conditions.

Because they might be influenced by multiple distinct factors, we analyze participants' perceptions of trust, security, and privacy by fitting linear regression models. For each regression analysis, we consider a set of candidate models and select the model with the lowest Akaike Information Criterion (AIC) [10]. We consider candidate models consisting of the "condition" (indicating the particular text disclosure, icon, or animation tested) plus every possible combination of optional factors. Required factors, optional factors, and corresponding baseline values are described in Table 1.

We present the outcomes of our regressions in tables where each row contains a factor and the corresponding change of the analyzed outcome in relation to the baseline of the given factor. Linear regression models measure change from baseline factors with a coefficient (*Coef.*) of zero for the value of the outcome. For each factor of a model, we also list a 95% confidence interval (*C.I.*) and a p-value indicating statistical significance. Also, we highlight p-values below $\alpha = 0.05$ with an asterisk (*).

We analyzed all free-text responses in an open-coding process [13, 60]. Two researchers iteratively developed a codebook [11], then used this initial codebook to code all free-text responses simultaneously, resolved coding conflicts, and incrementally updated the codebook until they were able to code open-ended questions without modifications to the codebook. The codebook remained stable once both researchers were satisfied that all important themes and concepts in the responses could be captured with the codes. Since the researchers resolved conflicts immediately as they emerged, we do not calculate inter-coder agreement [32].

3.2 Limitations

As with most self-reported online studies, our work has several limitations. In general, self-report studies may suffer from several biases, including over- and under-reporting, sampling bias, and social desirability bias. While we utilize self-report data, our central claims are not about the accuracy of respondents' answers to a given question, but rather about whether and how responses from different conditions differ from each other. Consequently, the threats to validity caused by those biases should apply equally across all conditions.

Conducting user studies on Amazon MTurk and Prolific is a widely used and accepted procedure for this type of research [39, 43]. However, MTurkers are known to be younger and more tech-savvy than the average population [43]. Additionally, our study focuses on the responses of U.S. Internet users, and thus, we can offer no insight into the generalizability of results for international participants.

Recently, the frequency of low data quality on MTurk has been increasing [31]. Therefore, we implemented a number of countermeasures (cf. Section 3.1). During data cleaning,

we identified several participants who did not pass our quality measures (Figure 1) and excluded them from further analysis.

We cannot guarantee that no participants were both registered MTurk and Prolific users and took more than one study, since there is no way to track people across both services. However, this is unlikely, since MTurk and Prolific target different geographic regions, and we conducted only the validation study on Prolific. Hence, we can guarantee that no participant took the same study twice.

Studies 2–5 tested a small set of different text disclosures, icons, and animations. While we based our designs on previous work and the results of our first study, we cannot guarantee that there are not other variants that work even better. Individual studies transpired in a somewhat isolated context, potentially missing certain effects of long-time exposure. We deliberately focused on multiple shorter studies, instead of one single in-depth, long-term study, to gather wider insights with different elements.

We showed our participants short screencasts (videos) in studies 2–5 instead of letting them use a real messaging application on their own devices. We aimed for high internal validity, so we wanted to ensure all participants would receive the same treatment. Comparable related work also worked with mockups instead of real applications for the same reason [18, 21]. While this experimental design results in lower external validity, we consider this tradeoff acceptable.

We decided to use a widely-deployed tool instead of a fictitious app mockup to study the challenges of visualizing E2E-encryption for an existing service provider and user base. We think our research provides valuable insights for a large set of users, although findings may not generalize to other E2E-encrypted messaging tools.

3.3 Ethical Considerations

We designed our studies with privacy in mind and followed best practices concerning data collection to ensure that we adhere to the German data- and privacy-protection laws aswell as the European General Data Protection Regulation. Our institution does not require a formal IRB, but we designed the study protocol based on a previous IRB approved study. All surveys started with a consent form to inform participants about the purpose of the study and about the data we would collect and store. The consent form also contained contact information to reach the PI in case of questions or concerns.

4 Use of Communication Tools (Study 1)

The main goal of Study 1 was to learn which communication tools our participants used and preferred for everyday and confidential conversations, as well as to learn about the decisions they made when using specific communication tools for particular conversations. In particular, we were interested in how many participants already used tools that provide E2E encryption by default for everyday conversations. We were especially interested in what fraction of them preferred less secure alternatives to E2E encrypted messengers for confidential conversations, and why. The questionnaire consisted of both closed- and open-ended questions. We followed the methodology described in Section 3 and developed the survey questionnaire in an iterative process, using pre-tests to improve the questionnaire, data quality, and determine appropriate compensation. We recruited 149 U.S.-based participants on MTurk.

4.1 Questionnaire Structure

We asked our participants to answer questions about their current use of communication tools for day-to-day and confidential conversations, as well as decisions they make when they choose one of the tools they have available for communicating with a single person or with groups of people. We decided to ask for specific tools or tool providers to glean insights into real behaviors and decision processes. We administered demographic questions at the end of the questionnaire to minimize stereotype bias [33, 53].

Past Tool Usage We asked participants which communication tools they have used in the last six months. The list of tools included the ten most popular tools in the U.S. [59]. We added iMessage, e-mail, and SMS to the list as popular messaging services that are pre-installed on many mobile devices by default. To better understand participants' choices and glean insights into their underlying mental models, we asked open-ended questions to explain their choices.

Security Assessments We asked participants to rate their perceived level of security when using personal e-mail, Facebook Messenger, WhatsApp, Snapchat, and SMS in the presence of different attackers. We chose these tools based on their popularity [2] and security properties (cf. Table 4 in the appendix).

Demographics We included several demographic questions about gender, age, ethnicity, education level, employment status, mobile device use, and the Security Behaviors Intentions Scale [19] for each participant. We aimed to assess whether demographic information would affect respondents' answers to the survey questionnaire. We also asked respondents for general feedback on the survey questionnaire.

4.2 Findings

We present both quantitative as well as qualitative results for the 149 valid respondents. The reporting of our findings focuses on actual tool usage in the past, insights into the perceptions, and decisions our participants made and their assessment of the security they think popular tools provide. Table 3 provides an overview of demographic characteristics of the participants in all studies.

Tool Usage Of the 149 participants in this study, the majority used regular e-mail (133; 89.26%), SMS/Text Messages (123; 82.55%) or the Facebook Messenger (114; 76.51%) that do not provide E2E encryption by default (cf. Figure 6). Only a few participants (7) reported having used PGP, S/MIME, or a provider supporting E2E encryption to secure e-mail conversations. Few participants (1) indicated prior use of Facebook's "Secret Conversation" feature. Overall, more than half of participants (80; 53.69%) reported use of an E2E-encrypted communication tool, with WhatsApp (47; 31.54%) being the most popular by far.

Tools that support E2E encryption as an optional feature, such as Facebook Messenger, Skype and Telegram (cf. Table 4), were also widely used (81.88%). However, only a few participants (9.02%) reported having used their E2E features.

While e-mail, SMS/text message, Facebook, WhatsApp, iMessage, and Skype are the most popular tools for both day-to-day and confidential conversations (cf. Figure 4 and Figure 5), a minority of participants (32; 21.48%) preferred none of the given tools for confidential conversations[1]. They only trusted non-digital forms of communication.

Even though they were users of E2E-encrypted communication tools for day-to-day conversations, many participants preferred e-mail and SMS for confidential conversations. Of the 80 participants who used E2E-encrypted tools for communication in general, the majority (50; 62.50%) preferred the use of insecure alternatives for confidential conversations. In particular, most (32; 68.09%) of the 47 WhatsApp users prefer less secure alternatives for confidential conversations.

Reasons for Using a Tool for Day-to-Day Conversations The main reason for people to use a certain communication tool for day-to-day conversation is ease of use (61.49%) followed by the availability of contacts in this tool (49.32%) and convenience (28.38%). One out of four (25.00%) participants also mentioned the delivery speed of text messages or instant messaging services and few (15.54%) mentioned the provided functionality. Some stated they are using a specific tool for a particular circle of people (11.49%) as mentioned by few participants: "*I belong to an online community for work and our main line of communication is through Facebook's messaging service.*" (P157), "*My husband uses Google hangouts too, and since I talk to him the most, this is the app I use most often.*" (P23), "*This is a group of family that has them, when I just need to relay info to that group I get on Telegram.*" (P27).

Few participants mentioned that they like a tool for storing a conversation history (5.41%), group chats (4.05%), message read info (5.40%) and disappearing messages (1.35%).

E-mail was an outlier as a preferred communication tool in many ways. Some participants (17.86% of e-mail users) prefer e-mail over other tools because they did not feel forced to reply to e-mails immediately:

"*It's more low key. There are no read receipts and you aren't expected to make a response immediately. You get to take your time.*" - P151.

E-mail has a professional reputation as it is often used in the workplace, which 16.06% of e-mail users noted. For 26.79% of e-mail users, a key reason to use e-mail is the support for large attachments and long text. This differs from all other tools, which are primarily instant-messaging services.

Reasons for Using a Tool for Confidential Conversations In two open-ended questions, we asked participants to elaborate on their preference for a specific tool for sensitive or confidential conversations and how they can tell that a specific tool keeps conversations confidential.

Almost half of participants (45.54%) mentioned a gut instinct that leads to a security belief as their main reason to prefer a specific tool for confidential conversations, e. g. "*I feel that it is safe.*" (P36).

A quarter of our participants (25.00%) assumed a tool to be confidential when they send messages directly to their intended contact and had their own name and the name(s) of the communication partner(s) being shown in the user interface. 16.96% mentioned access control and strong passwords as reasons to prefer a particular tool as mentioned by one participant: "*I have a secure E-mail that is guarded by a good strong password.*" (P123).

One out of four (26.79%) assumed a tool to be acceptable for confidential conversations because they thought it uses some form of encryption. However, 14.29% made wrong assumptions and thought encryption was being deployed on unencrypted channels (e. g., for SMS/text messages). Interestingly, only a few (8.04%) referenced "secret mode" or "secure chat" options in their decision.

6% of our participants also reported using SMS as a confidential channel because it is not an internet service: "*It is sent from me to another person, not on the internet.*" (P31) and "*It feels off the grid, away from the dangers of the internet.*" (P66)

For a few (3.57%), visual indicators like colors or icons earned trust even if they did not directly relate to security or privacy e. g. "*If the message is blue it should be encrypted.*" (P148). In the iOS messenger, a blue message indicates that a message was sent via iMessage and a green message indicates that it was sent as a Text Message.

Few participants mentioned self-destructing and disappearing messages (4.46%), as in SnapChat, or the ability to delete messages manually (3.57%), as offered in WhatsApp, as influencing their preference:

"*I know that gmail for example encrypts messages and I trust google to be safe.*" (P77)

At the same time, half of the participants (50%) could not report specific reasons for their trust in a particular tool.

> Key Insights: Tool Usage, Decisions and Security Beliefs.
>
> - E-mail and SMS/text messages are the most popular tools for both day-to-day and confidential conversations.

[1] None is an exclusive option and deselected the other fields.

- 53.69% of participants use a communication tool with E2E encryption enabled by default.
- 62.50% of participants who use E2E-encrypted tools prefer less secure alternatives for confidential conversations.
- Participants reported a gut instinct that made them believe a tool to be secure.

5 Visualizing Encryption (Study 2–4)

Both previous work and the findings of our first study illustrate that the situation around E2E-encrypted communication tools is complicated. Many users will avoid installing a new, more secure messaging tool [2] only because it provides better security [51, 67]. Instead, most users only consider messaging tools if their contacts (i. e. friends, family, and colleagues) also use the tools [14]. Additionally, previous work [14, 68], and our first study show many people suffer from misunderstandings and misconceptions of encryption.

Instead of propagating the more widespread use of such niche tools or working on correcting users' misunderstandings and misconceptions alone, we followed a different route. Depending on geographic region, between half of users (cf. Section 4) and 90% [2] of users *already* have tools that support E2E encryption by default, with WhatsApp being the most popular. However, our findings (cf. Section 4) suggest that many users are not aware of these security properties. More than half of our participants who use WhatsApp prefer less secure alternatives such as e-mail or SMS/text messages for confidential conversations. Therefore, the remainder of our studies investigate how visualizing encryption impacts perceptions of E2E messaging security.

While the results of our first study (cf. Section 4) and previous work [3, 4, 14, 18, 21, 68] uncover a wide range of root causes for misconceptions about the security of messengers and insecure behaviour, only some of them can be addressed in the design of a communication tool. For example, we identified that trusting a company or decades of positive experiences were both root causes for misconceptions. However, these can hardly be addressed in the design of a communication tool. In contrast, there are promising candidates that can directly be implemented in the user interface of a communication tool (cf. Section 3). In this section, we describe multiple online studies we conducted with the goal to investigate the impact of different encryption disclosures, icons and animations on perceived trust, security, privacy, usability, satisfaction and self-reported likeliness to use the re-designed communication service for sensitive messages.

5.1 Experiment Design

To study visualizations of encryption using text disclosures, icons, and animations, we conducted four between-groups online experiments with WhatsApp users recruited on MTurk or Prolific. Each study follows the procedure we outline below.

5.1.1 Screencasts

To study the impact of different encryption visualizations on usability, perceived trust, security, privacy, satisfaction and tool preference, we decided to show participants a screencast of a ficticious WhatsApp update [2].

Using a screencast instead of static mockup images allowed us to study both static and dynamic encryption visualizations (cf. Section 3) and include a scripted conversation to provide more context for our participants[3]. We constructed the messages this way because it mimics a realistic personal conversation and credit card information is generally perceived as confidential and worth protecting.

To mimic WhatsApp as closely as possible, we forked the Android version of the Signal mobile app and adapted the user interface respectively by changing colors, typefaces, buttons and other user interface properties.

5.1.2 Questionnaire Structure

The survey questionnaire in this study was developed through an iterative process (cf. Section 4) and included the attention checks mentioned in Section 3. Completion of the survey took 10 minutes on average and we paid participants $1.7.

Usability, Trust, Security, Privacy and Satisfaction We asked participants to answer usability, perceived trust, security and privacy and satisfaction questions. For usability, we asked participants the two items UMUX lite scale [34]. Based on prior work [44], we built a 10-item scale of perceived trust, security and privacy (cf. Appendix A.2). Finally, we asked participants to fill out the net promoter score [28] to measure how much they liked the encryption visualization.

Tool Preference We showed participants a list of the most popular communication tools from our first study (cf. Section 4) including the new ficticious WhatsApp version and asked them which tool they would prefer for both day-to-day and confidential conversations. To prevent lock-in obstacles as found in [14], we told all participants to assume that all communication partners have all tools installed. We aimed to assess whether our conditions had an effect on the participants' choice.

Demographics We asked our respondents the same demographic questions as in the questionnaire in Section 4. Table 3 provides an overview of demographic characteristics of the participants in the studies in this section.

5.2 Text Disclosures (Study 2)

Based on the disclosures in current tools that support E2E-encrypted communication (cf. Table 4) and the results of our first study (cf. Section 4), we created six different disclosures out of the terms "secret", "private", "encrypted", "secure"

[2] cf. Appendix A.6 for the video introduction
[3] cf. Appendix A.4 for the conversation

	Factor	Coef.	C.I.	*p*-value
Disclosure	"*Messages to this chat are now . . .*"			
	"*. . . private*"	0.27	[0.21, 1.07]	0.392
	"*. . . secret*"	-0.49	[-1.09, 0.11]	0.111
	"*. . . secure*"	0.06	[-0.53, 0.65]	0.845
	"*. . . encrypted*"	0.68	[0.09, 1.28]	0.030 *
	"*. . . end-to-end encrypted*"	-0.09	[-0.69, 0.51]	0.768
	"*. . . secured with end-to-end encryption*"	0.41	[-0.19, 1.01]	0.182
Icon	Icon (Baseline: Control):			
	Envelope	-0.47	[0.92, 1.69]	0.105
	Lock	-0.49	[-1.09, 0.11]	0.089
	Shield	-0.70	[-1.27, -0.14]	0.014 *
	CS Education	-0.51	[-0.97, -0.05]	0.029 *
Animation	Animation (Baseline: Control):			
	Disappearing Messages	0.08	[-0.34, 0.51]	0.707
	Encryption/Decryption	-0.01	[-0.40, 0.38]	0.969
	Progress Circle	0.25	[-0.14, 0.66]	0.210
	Age	-0.01	[-0.14, 0.65]	0.119
Validation	Animation (Baseline: Control without Disclosure):			
	Control	0.45	[0.03, 0.86]	0.034 *
	Disappearing Messages	0.40	[0.01, 0.79]	0.043 *
	Encryption/Decryption	0.43	[0.04, 0.81]	0.030 *
	Progress Circle	0.71	[0.33, 1.10]	< 0.001 *

Table 2: Results of the linear regression model examining whether different texts, icons and animations have an effect on the trust, security and privacy score in relation to a control baseline. Note the additional "Control" variable in the last study, due to "Control Without Disclosure" being the baseline. See Table 1 for further details.

and "end-to-end encrypted." In a between-groups design, we randomly assigned participants to one of the following conditions:

1. Control: "*blank*"
2. Encrypted: "*Messages to this chat are now **encrypted**.*"
3. E2E-Encrypted: "*Messages to this chat are now **end-to-end encrypted**.*"
4. Private: "*Messages to this chat are now **private**.*"
5. Secure: "*Messages to this chat are now **secure**.*"
6. Secure & E2E: "*Messages to this chat are now **secured** with **end-to-end encryption**.*"
7. Secret: "*Messages to this chat are now **secret**.*"

To make sure participants read the text of each disclosure, we showed them a screencast in fullscreen before entering the conversation for seven seconds including the respective disclosure. Overall, we recruited 196 valid participants on MTurk for whom we report findings below.

Findings We were specifically interested in the participants' opinions on usability and their perceptions of trust, security, and privacy.

As a usability metric we compared the distribution of UMUX Lite answer categories between our conditions. We found no apparent differences between the conditions (Q1: Pearson's $\chi^2 = 1.56$, p-value $= 1$; Q2: Pearson's $\chi^2 = 1.22$, p-value $= 1$), suggesting no observable effect (positive or negative) on the perceived usability of the different disclosures.

To better investigate how the different conditions affect participants' perception of trust, privacy, and security, we introduced a combined score based on their answers to our set of 10 likert-item questions. For each participant, the score consists of the average of all 10 likert-item questions mapped to numerical values, e.g., between -2 (Strongly Disagree) and +2 (Strongly Agree). For these scores, we considered a set of linear regression models consisting of the conditions as required factor and all combinations of optional factors listed in Table 1 and selected the model with the lowest AIC.

The final model (see Table 2) shows that the "encrypted" condition is significant with an overall positive coefficient of about 0.7 score points compared to the control baseline. This suggests significantly higher scores for the "*Messages to this chat are now encrypted*" disclosure compared to the blank control, which is in line with previous research by Distler et al. [17]. Participants seemed to prefer the encryption text, likely due to not fully understanding the term "end-to-end," or regarding it as a subset (i.e., less secure) of being "just" encrypted.

For the net promoter score, we found that no condition dominates any other (Kruskal-Wallis $H = 7.88$, p-value $= 0.24$). Based on these results, we proceeded with the "encrypted" text for our subsequent disclosure.

> Key Insights: Disclosures.
>
> - Participants felt most secure and private within the "encrypted" disclosure condition.
> - The different disclosures did not have a significant impact on usability and satisfaction.

5.3 Icons (Study 3)

Next, we investigated three different icons. We chose a lock, a shield, and an envelope (cf. Figure 2) based on their typical usage in security and privacy contexts [7, 38], previous work in the field of security indicators [23, 54], and results from our first study (cf. Section 4). Together with the best-performing text disclosure from the previous study *"Messages to this chat are now encrypted,"* we showed participants one of the icons before the communication partners in the screencast entered the conversation.

(a) Envelope (b) Lock (c) Shield

Figure 2: Designs used in the encryption icons study.

We showed participants a screencast including the "encrypted" disclosure from the previous study and the respective

encryption icon. All screencasts lasted 94 seconds. A total of 86 WhatsApp users participated in this study. Findings reported below are limited to these valid participants.

Findings For the UMUX Lite questionnaire we found no significant differences across conditions (Q1: Pearson's $\chi^2 = 0.54$, p-value $= 1$; Q2: Pearson's $\chi^2 = 0.40$, p-value $= 1$).

Our set of linear regression models for the overall score included the icon condition as required factor and again all combinations of optional factors (cf. Table 1). To our surprise, the final model (See Table 2) shows that all three icon conditions are worse than the baseline by at least 0.47 score points. The shield condition is significantly worse by 0.7 score points. In addition, the optional computer science education factor is significant with a negative coefficient in the model. This is in line with previous work [23, 54] and additional evidence for the very limited effect of security icons on perceived trust, security and privacy.

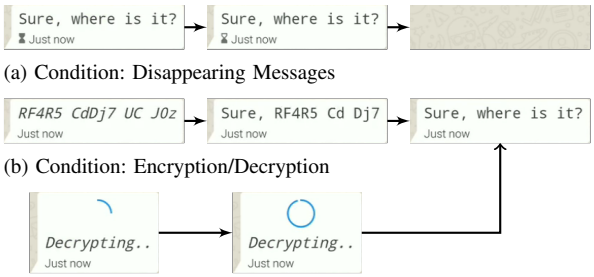
For the net promoter score, we found that no condition dominates any other (Kruskal-Wallis $H = 5.68$, p-value $= 0.128$). We chose to proceed to the next study using the control (no icon) due to the negative coefficients of all other conditions.

> Key Insights: Security Icons.
>
> - We found a negative effect of security icons on perceived trust, security and privacy by at least 0.47 score points compared to the baseline.
> - Participants with a computer science background particularly disliked the security icons we investigated, resulting in 0.51 less score points compared to participants without that background.
> - The security icons had no impact on usability and satisfaction.

5.4 Animations (Study 4)

In addition to text disclosures and encryption icons, previous work [18, 21] and the results of our first study (cf. Section 4) suggest the use of animations of the encryption process to convey that a conversation is secure. We identified three different encryption animations: (i) Distler et al. [18] used a progress circle for an e-voting app; (ii) Fahl et al. [21] studied dynamic encryption and decryption animations to protect Facebook messages; and (iii) participants in Study 1 reported feeling particularly secure with disappearing messages on apps like Snapchat. Although disappearing messages are not technically connected to E2E-encryption, we included them due to their contribution to perceived messaging security identified in previous work by Roesner et al. [45] and participants' comments in Study 1. One participant for example said it was security relevant "*Because the conversation deletes right after I read it.*" (P9) and another said "*... once you open it, it's gone forever afterward.*" (P14)

We implemented those three animations (cf. Figure 3). In contrast to Studies 2–3, we applied the dynamic encryption animations to the screencast conversation's messages, rather than as a fullscreen hint before entering a conversation. We



(a) Condition: Disappearing Messages

(b) Condition: Encryption/Decryption

(c) Condition: Progress Circle

Figure 3: Conditions in the security animations study.

pre-tested the animation duration with 20 MTurkers. Initially, we showed the animation for three seconds. Based on UMUX Lite and qualitative feedback, we gradually reduced the duration to one second. Based on the results from the previous studies, each condition included one of the encryption animations and a small text hint with the "Encrypted" disclosure. Based on our Study 3 results, we did not use an icon in this study.

Overall, we recruited 107 valid participants on MTurk for whom we report findings below.

Findings We found no impact of the different animation conditions on the UMUX Lite questionnaire (Q1: Pearson's $\chi^2 = 0.37$, p-value $= 1$; Q2: Pearson's $\chi^2 = 0.42$, p-value $= 1$). The final linear regression model includes a somewhat increased, but not significant, coefficient (0.25) for the "Progress Circle Animation" condition compared to the baseline, and almost non-existent positive and negative effects (0.08 and -0.01) for the other two animations (cf. Table 2). For the net promoter score of the animation conditions, we found again, as in the other two studies, that no condition dominates any other (Kruskal-Wallis $H = 1.63$, p-value $= 0.654$).

> Key Insights: Security Animations.
>
> - No factors were significant. The progress circle animation had a weak positive effect on perceived trust, security and privacy.
> - Security animations did not impact usability or satisfaction.

6 Validation (Study 5)

Due to Study 4's inconclusive findings regarding the effect of animations (cf. Section 5.4), we validated our overall findings in a fifth study. Motivated by the lower data quality encountered in Study 4 compared to Studies 1–3, as well as the increasing difficulty recruiting WhatsApp users on MTurk, we switched to the Prolific recruitment platform for Study 5. Prolific provides fine-grained participant demographics, so we could directly target participants with messenger experience without requiring a qualification task. Since Prolific's pool of US workers who use WhatsApp was small (<500), we included UK participants, increasing the participant pool by 6,000. On Prolific, we recruited 145 participants from the UK and US, paying £2.70. In addition to changing recruitment

platforms, we decided to delve into the root causes of why visualizing encryption appeared to have a limited impact on user perceptions. For this, we dropped both the UMUX Lite and net promoter score as we observed no significant differences for them in our previous studies. We also added open-ended followups to each Likert question to gain deeper insights into participants' opinions. Additionally, we added Likert questions that focused on what facets suggest that messages are being sent securely. To eliminate participants' perceptions of WhatsApp as a major factor guiding their perceptions, we also changed the messenger name to "Erebus." As this study was intended as validation, we especially focused on the "encryption" text from our first survey by including a new control that displayed no text at all ("Control without Text"). We retained the previous control condition to compare with the previous studies.

Findings As for the previous surveys, we generated a linear regression model listed in Table 2. The final regression models have significant non-zero coefficients for all included variables relative to our new control (not mentioning encryption at all). Going by coefficient, the Progress animation performs best compared to the baseline (0.71), followed by Encryption/Decryption (0.43) and Disappearing (0.40). Even the control condition from the previous surveys ("Control") shows a significant coefficient compared to the newly introduced baseline "Control without Text". This suggests a significant effect of the "encryption" text.

In addition to the regression analysis, we evaluated the open-ended questions to gain insight into the limited impact of encryption visualization. We report findings below.

Observing Animations In an open-ended question, we asked participants what they observed happening (if anything) when messages were sent or received, as well as what this indicated. Almost all participants described the animations we showed them ($> 90\%$ in each condition). 60 participants (41.38%) wrote that the animation indicated an increased level of security. Hence, we can eliminate the possibility of participants ignoring the animations as the reason perceptions did not vary significantly across conditions.

Identifying Security In an open-ended question, we asked participants how they determine, in general, that a messaging app sends messages securely. The most prominent indicator for security was the *reputation* (48, 33.10%) of the service provider, followed by the mention of *encryption* (42, 28.96%).

> "*Honestly, I guess I just trust in the brand that it's safe. I do this through the popularity, good press and confidence in their service.*" - P18

The relatively similar relevance of *encryption* and *reputation* for perceived security also explains the limited impact of the presence of encryption on perceptions of security.

Identifying Encryption Given that encryption is an important security mechanism, we asked participants to detail how they identify the presence of encryption in a messaging app. Most participants report relying on textual information in the form of *disclosures* (40, 27.59%) or an app's *feature list* (11, 7.59%) mentioning encryption. However, 42 participants (28.97%) said they would not know how to recognize encryption's presence. Very few mentioned visual indicators. For example, 5 (3.44%) mentioned observing a *delay during sending*, 3 (2.07%) mentioned messages disappearing, and 2 (1.38%) mentioned seeing messages be *scrambled*. These explanations are consistent with our regression analyses (cf. Table 2), highlighting the limited effect of visualizations.

> Key Insights: Validation.
>
> - Study 5 confirmed the findings of Studies 2–4.
> - Visualizing encryption in any way, even a simple text disclosure, improves perceptions compared to not mentioning it at all.
> - Most participants saw the animations and felt they communicated "security", yet this did not change their perceptions any more than a text disclosure did.
> - An app's reputation greatly impacts perceptions.

7 Discussion

In our first of five studies, we investigated why participants use particular messaging tools, validating a prior finding [3] that many users mistakenly think SMS and e-mail are more secure than E2E-encrypted messengers. Based on these initial findings, we aimed to improve the visibility of E2E encryption in a messaging app. Across the four subsequent studies, we compared six different text disclosures, three different icons, and three different animations of the encryption process.

Impact of Encryption Visualization While investigating the impact of different visualizations of encryption, we were surprised to find that the simple "encrypt" disclosure outperformed most others (aside from the progress circle) in terms of perceived trust, security, and privacy. As expected, however, all disclosures performed better than the baseline of having no disclosure at all. We were also surprised to see that security icons had a negative effect, rather than increasing perceptions of trust, security, and privacy. This negative effect was particularly distinct for people with a CS background.

Previous work suggested that encryption visualizations might positively impact perceived trust, security, and privacy [18, 21, 48]. Those suggestions were based primarily on qualitative data. Our studies, which combined quantitative and qualitative data, reached somewhat different conclusions. Based only on the qualitative data we collected, one might have reached conclusions similar to those of prior work. For example, as reported in Section 6, nearly half of participants indicated that the animations of encryption indicated an increased level of security. In contrast, our quantitative analyses indicated that these different animations did not have a significantly different impact on perceptions of the trust, security, and privacy of E2E-encrypted messaging tools than a straightforward text disclosure that the conversation in encrypted, which is what many secure messaging apps currently display.

These findings call into question the magnitude and applicability of the effects reported in prior work.

Our findings suggest that highlighting the use of encryption in basic ways (e.g., "*Messages to this chat are now encrypted*") significantly increases perceived security, privacy and trust in messaging applications. That is, having *any* visualization of encryption outperformed the control in our validation study of not calling attention to the use of encryption at all. However, richer visualizations of encryption involving icons or animations seem to have only a limited additional effect. Although we did not observe these richer visualizations of encryption to significantly impact user perceptions and satisfaction in a positive direction compared to basic text disclosures, we also did not observe a negative effect.

8 Recommendations

Given the promise of rich visualizations of encryption reported in prior work, this finding is disappointing, as it suggests that simple modifications of messaging apps' UIs are unlikely to help users better assess apps' security and privacy. Despite the use of multiple design proposals from previous work, we could not find a significant improvement (Sec. 7).

Our qualitative results imply that instead of investing more effort into studying richer visualizations of encryption, focusing on the following aspects is potentially more promising. We make recommendations for both providers of E2E encrypted communication tools and usable security researchers.

8.1 Tool Providers

Trust in Company As we have seen in the qualitative answers in the tool usage and validation study (Sec. 4.2, 6), participants report that they trust the brand and that the company would keep their data secure. Tool providers could focus on generally improving trust in the brand.

Convenience Several participants mentioned using a specific app to communicate with their peer groups that decided on that app (Sec. 4.2). Introducing an app or feature that is not compatible with their peer groups leads to them switching back to another channel. Tool providers should make sure that E2E encryption features do not lead to inconveniences for their users.

Functionality Our participants also mentioned that they switched the tools when a messenger did not support a required feature, for example with large attachments that they send via mail (Sec. 4.2). That indicates that a full feature set is required to avoid people switching to insecure channels. Making E2E encryption available in communication tools should not limit existing functionality.

8.2 Usable Security Research

Correcting Mental Models Our participants showed a number of incorrect mental models, most strikingly: Around 25% of our participants assumed that their conversations are free of eavesdroppers if the user interface shows only the names of their intended communication partner(s) (Sec. 4.2) and show a lack of understanding of man-in-the-middle attacker capabilities. Also, 14.29% of participants falsely assumed channels that are generally not encrypted by default (e.g., SMS) to be encrypted (Sec. 4.2). These misconceptions likely impact the usage of secure and private messengers significantly. Addressing them better should be a major goal for our community.

Technical Background As seen in our regression for Study 3 (Table 2), participants with a technical background tended to rate trust in security indicators lower. Investigating factors that contribute to this perception and provide improvements for these factors (e.g., increase company transparency) could help address concerns unique to that demographic.

9 Conclusion

We studied whether making a messaging app's E2E encryption more visible improves perceptions of trust, security, and privacy. To that end, we conducted five online studies with 683 total participants, including a summative validation study.

While participants felt most secure and private within the "encrypted" text disclosure condition, the different text disclosures did not have a significant impact on usability and app satisfaction. We observed a surprising negative effect of security icons on perceived trust, security, and privacy. When focusing on animations, none of the factors was statistically significant, though we identified a weak positive effect for the progress circle animation on perceived trust, security, and privacy. The animations had no impact on usability and satisfaction. We confirmed these key findings in a final summative study, validating that visualizing encryption in any way, even a simple text disclosure, improves perceptions compared to not mentioning encryption at all. Most participants saw the animations, the richest and most novel aspect of our investigation, and reported qualitatively they communicated "security." However, quantitative perceptions of trust, security, and privacy did not differ significantly compared to a text disclosure.

In our first study, we replicated the finding of prior work that a non-trivial fraction of users mistakenly believes SMS and e-mail to be more secure than E2E-encrypted messengers. While we had hypothesized that richly visualizing the process of encryption would emphasize E2E-encrypted messaging apps' security properties and combat this misconception, our results suggest that the existing practice of disclosing the use of encryption in a straightforward text disclosure may be sufficient if the text disclosure is displayed prominently.

References

[1] Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. More than Smart Speakers: Security and Privacy Perceptions of Smart Home Personal Assistants. In

Proc. 15th Symposium on Usable Privacy and Security (SOUPS'19). USENIX Association, 2019.

[2] Ruba Abu-Salma, Kat Krol, Simon Parkin, Victoria Koh, Kevin Kwan, Jazib Mahboob, Zahra Traboulsi, and M Angela Sasse. The Security Blanket of the Chat World: An Analytic Evaluation and a User Study of Telegram. In *Proc. 2nd European Workshop on Usable Security (EuroUSEC'17)*. The Internet Society, 2017.

[3] Ruba Abu-Salma, Elissa M. Redmiles, Blase Ur, and Miranda Wei. Exploring user mental models of end-to-end encrypted communication tools. In *Proc. 8th USENIX Workshop on Free and Open Communications on the Internet (FOCI'18)*. USENIX Association, 2018.

[4] Ruba Abu-Salma, M. Angela Sasse, Joseph Bonneau, Anastasia Danilova, Alena Naiakshina, and Matthew Smith. Obstacles to the Adoption of Secure Communication Tools. In *Proc. 38th IEEE Symposium on Security and Privacy (SP'17)*. IEEE, 2017.

[5] Omer Akgul, Wei Bai, Shruti Das, and Michelle L. Mazurek. Evaluating In-Workflow Messages for Improving Mental Models of End-to-End Encryption. In *Proc. 30th Usenix Security Symposium (SEC'21)*. USENIX Association, 2021.

[6] Apple Inc. iMessage - Learn more about Messages. https://support.apple.com/explore/messages, 2019.

[7] Apple Inc. Human Interface Guidelines - iOS Design Themes. https://developer.apple.com/ios/human-interface-guidelines/overview/themes/, 2020.

[8] Erinn Atwater, Cecylia Bocovich, Urs Hengartner, Ed Lank, and Ian Goldberg. Leading Johnny to Water: Designing for Usability and Trust. In *Proc. 11th Symposium on Usable Privacy and Security (SOUPS'15)*. USENIX Association, 2015.

[9] Wei Bai, Michael Pearson, Patrick Gage Kelley, and Michelle L. Mazurek. Improving Non-Experts' Understanding of End-to-End Encryption: An Exploratory Study. In *Proc. 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 2020.

[10] K. P. Burnham. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2):261–304, 2004.

[11] Kathy Charmaz. *Constructing Grounded Theory*. SAGE Publications, 2014.

[12] Sandy Clark, Travis Goodspeed, Perry Metzger, Zachary Wasserman, Kevin Xu, and Matt Blaze. Why (Special Agent) Johnny (Still) Can't Encrypt: A Security Analysis of the APCO Project 25 Two-Way Radio System. In *Proc. 20th Usenix Security Symposium (SEC'11)*. USENIX Association, 2011.

[13] Juliet Corbin and Anselm Strauss. Grounded theory research: Procedures, canons and evaluative criteria. *Zeitschrift für Soziologie*, 19(6):418–427, 1990.

[14] Alexander De Luca, Sauvik Das, Martin Ortlieb, Iulia Ion, and Ben Laurie. Expert and Non-Expert Attitudes towards (Secure) Instant Messaging. In *Proc. 12th Symposium on Usable Privacy and Security (SOUPS'16)*. USENIX Association, 2016.

[15] Sergej Dechand, Dominik Schürmann, Karoline Busse, Yasemin Acar, Sascha Fahl, and Matthew Smith. An Empirical Study of Textual Key-Fingerprint Representations. In *Proc. 25th Usenix Security Symposium (SEC'16)*. USENIX Association, 2016.

[16] Albesë Demjaha, Jonathan Spring, Ingolf Becker, Simon Parkin, and M. Angela Sasse. Metaphors considered harmful? An exploratory study of the effectiveness of functional metaphors for end-to-end encryption. In *Proc. Workshop on Usable Security (USEC'18)*. The Internet Society, 2018.

[17] Verena Distler, Carine Lallemand, and Vincent Koenig. Making Encryption Feel Secure: Investigating how Descriptions of Encryption Impact Perceived Security. In *Proc. 5th European Workshop on Usable Security (EuroUSEC'20)*. IEEE, 2020.

[18] Verena Distler, Marie-Laure Zollinger, Carine Lallemand, Peter B. Rønne, Peter Y. A. Ryan, and Vincent Koenig. Security - Visible, Yet Unseen? In *Proc. CHI Conference on Human Factors in Computing Systems (CHI'19)*. ACM, 2019.

[19] Serge Egelman and Eyal Péer. Scaling the Security Wall: Developing a Security Behavior Intentions Scale (SeBIS). In *Proc. 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15)*. ACM, 2015.

[20] Ksenia Ermoshina, Francesca Musiani, and Harry Halpin. End-to-End Encrypted Messaging Protocols: An Overview. In *Proc. 6th International Conference on Internet Science (INSCI'19)*. Springer, 2016.

[21] Sascha Fahl, Marian Harbach, Thomas Muders, Matthew Smith, and Uwe Sander. Helping Johnny 2.0 to Encrypt His Facebook Conversations. In *Proc. 8th Symposium on Usable Privacy and Security (SOUPS'12)*. ACM, 2012.

[22] Adrienne Porter Felt, Alex Ainslie, Robert W. Reeder, Sunny Consolvo, Somas Thyagaraja, Alan Bettes, Helen Harris, and Jeff Grimes. Improving SSL Warnings: Comprehension and Adherence. In *Proc. 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15)*. ACM, 2015.

[23] Adrienne Porter Felt, Robert W. Reeder, Alex Ainslie, Helen Harris, Max Walker, Christopher Thompson, Mustafa Embre Acer, Elisabeth Morant, and Sunny Consolvo. Rethinking Connection Security Indicators. In *Proc. 12th Symposium on Usable Privacy and Security (SOUPS'16)*. USENIX Association, 2016.

[24] Ronald P Fisher and R Edward Geiselman. *Memory-Enhancing Techniques for Investigative Interviewing: The Cognitive Interview*. Charles C Thomas Publisher, 1992.

[25] Simson L. Garfinkel, David Margrave, Jeffrey I. Schiller, Erik Nordlander, and Robert C. Miller. How to Make Secure Email Easier To Use. In *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI'05)*. ACM, 2005.

[26] Simson L. Garfinkel and Robert C. Miller. Johnny 2: A User Test of Key Continuity Management with S/MIME and Outlook Express. In *Proc. 1st Symposium on Usable Privacy and Security (SOUPS'05)*. ACM, 2005.

[27] Nina Gerber, Verena Zimmermann, Birgit Henhapl, Sinem Emeröz, and Melanie Volkamer. Finally Johnny Can Encrypt: But Does This Make Him Feel More Secure? In *Proc. 13th International Conference on Availability, Reliability and Security (ARES'18)*. ACM, 2018.

[28] D. F. Hamilton, J. V. Lane, P. Gaston, J. T. Patton, D. J. MacDonald, A. H. R. W. Simpson, and C. R. Howie. Assessing treatment outcomes using a single question: the net promoter score. *The Bone & Joint Journal*, 96(5):622–628, 2014.

[29] Julia Hanson, Miranda Wei, Sophie Veys, Matthew Kugler, Lior Strahilevitz, and Blase Ur. Taking Data Out of Context to Hyper-Personalize Ads: Crowdworkers' Privacy Perceptions and Decisions to Disclose Private Information. In *Proc. CHI Conference on Human Factors in Computing Systems (CHI'20)*. ACM, 2020.

[30] David J. Hauser and Norbert Schwarz. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, 48(1):400–407, 2016.

[31] Ryan Kennedy, Scott Clifford, Tyler Burleigh, Ryan Jewell, and Philip Waggoner. The Shape of and Solutions to the MTurk Quality Crisis. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3272468, 2018.

[32] Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology (2nd ed.)*. SAGE Publications, 2004.

[33] Jon A. Krosnick and Stanley Presser. *Question and Questionnaire Design*, pages 263–314. Emerald Publishing, 2010.

[34] James R. Lewis, Brian Utesch, and Deborah E. Maher. UMUX-LITE: when there's no time for the SUS. In *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*. ACM, 2013.

[35] Max-Emanuel Maurer, Alexander De Luca, and Tobias Stockinger. Shining Chrome: Using Web Browser Personas to Enhance SSL Certificate Visualization. In *Proc. 13th IFIP TC 13 International Conference on Human-Computer Interaction (INTERACT '11)*. Springer, 2011.

[36] Micah Lee. Edward snowden explains how to reclaim your privacy. https://theintercept.com/2015/11/12/edward-snowden-explains-how-to-reclaim-your-privacy/, 2015.

[37] Mainack Mondal, Günce Su Yilmaz, Noah Hirsch, Mohammad Taha Khan, Michael Tang, Christopher Tran, Chris Kanich, Blase Ur, and Elena Zheleva. Moving Beyond Set-It-And-Forget-It Privacy Settings on Social Media. In *Proc. 26th ACM Conference on Computer and Communication Security (CCS'19)*. ACM, 2019.

[38] Jakob Nielsen. Enhancing the Explanatory Power of Usability Heuristics. In *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI'94)*. ACM, 1994.

[39] Stefan Palan and Christian Schitter. Prolific.ac — A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

[40] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153 – 163, 2017.

[41] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, 46(4):1023–1031, 2014.

[42] Prolific. Prolific | Online participant recruitment for surveys and market research. https://prolific.co/, 2020.

[43] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples. In *Proc. 40th IEEE Symposium on Security and Privacy (SP'19)*. IEEE, 2019.

[44] Juan Carlos Roca, Juan José García, and Juan José de la Vega. The importance of perceived trust, security and privacy in online trading systems. *Information Management & Computer Security*, 17(2):96–113, 2009.

[45] Franziska Roesner, Brian T Gill, and Tadayoshi Kohno. Sex, Lies, or Kittens? Investigating the Use of Snapchat's Self-Destructing Messages. In *Proc. 18th International Conference on Financial Cryptography and Data Security (FC'14)*. Springer, 2014.

[46] Christoph Rottermanner, Peter Kieseberg, Markus Huber, Martin Schmiedecker, and Sebastian Schrittwieser. Privacy and Data Protection in Smartphone Messengers. In *Proc. 17th International Conference on Information Integration and Web-based Applications & Services (iiWAS'15)*. ACM, 2015.

[47] Scott Ruoti, Jeff Andersen, Scott Heidbrink, Mark O'Neill, Elham Vaziripour, Justin Wu, Daniel Zappala, and Kent Seamons. "We're on the Same Page": A Usability Study of Secure Email Using Pairs of Novice Users. In *Proc. CHI Conference on Human Factors in Computing Systems (CHI'16)*. ACM, 2016.

[48] Scott Ruoti, Jeff Andersen, Travis Hendershot, Daniel Zappala, and Kent E. Seamons. Private Webmail 2.0: Simple and Easy-to-Use Secure Email. In *Proc. 29th Annual Symposium on User Interface Software and Technology (UIST'16)*. ACM, 2016.

[49] Scott Ruoti, Jeff Andersen, Daniel Zappala, and Kent E. Seamons. Why Johnny Still, Still Can't Encrypt: Evaluating the Usability of a Modern PGP Client. *ArXiv e-prints*, 2016.

[50] Scott Ruoti, Nathan Kim, Ben Burgon, Timothy Van Der Horst, and Kent Seamons. Confused Johnny: When Automatic Encryption Leads to Confusion and Mistakes. In *Proc. 9th Symposium on Usable Privacy and Security (SOUPS'13)*. ACM, 2013.

[51] M. A. Sasse, S. Brostoff, and D. Weirich. Transforming the 'Weakest Link' — a Human/Computer Interaction Approach to Usable and Effective Security. *BT Technology Journal*, 19(3):122–131, 2001.

[52] J. Schaad, B. Ramsdell, and S. Turner. Secure/Multipurpose Internet Mail Extensions (S/MIME) Version 4.0 Message Specification. https://tools.ietf.org/html/rfc8551, April 2019.

[53] Nora Cate Schaeffer and Stanley Presser. The Science of Asking Questions. *Annual Review of Sociology*, 29(1):65–88, 2003.

[54] Stuart E Schechter, Rachna Dhamija, Andy Ozment, and Ian Fischer. The Emperor's New Security Indicators. In *Proc. 28th IEEE Symposium on Security and Privacy (SP'07)*. IEEE, 2007.

[55] Svenja Schröder, Markus Huber, David Wind, and Christoph Rottermanner. When SIGNAL hits the Fan: On the Usability and Security of State-of-the-Art Secure Mobile Messaging. In *Proc. 1st European Workshop on Usable Security (EuroUSEC'16)*. The Internet Society, 2016.

[56] Signal, a 501c3 nonprofit. Signal Messenger. https://www.signal.org, 2019.

[57] Jennifer Sobey, Robert Biddle, Paul C. van Oorschot, and Andrew S. Patrick. Exploring User Reactions to New Browser Cues for Extended Validation Certificates. In *Proc. 13th European Symposium on Research in Computer Security (ESORICS'08)*. Springer, 2008.

[58] Andreas Sotirakopoulos, Kirstie Hawkey, and Konstantin Beznosov. On the challenges in usable security lab studies: Lessons learned from replicating a study on ssl warnings. In *Proc. 7th Symposium on Usable Privacy and Security (SOUPS'11)*. ACM, 2011.

[59] Statista Inc. Most popular mobile messaging apps in the United States as of September 2019, by monthly active users. https://www.statista.com/statistics/350461/mobile-messenger-app-usage-usa, 2019.

[60] Anselm Strauss and Juliet M Corbin. *Grounded theory in practice*. SAGE Publications, 1997.

[61] Joshua Sunshine, Serge Egelman, Hazim Almuhimedi, Neha Atri, and Lorrie Faith Cranor. Crying Wolf: An Empirical Study of SSL Warning Effectiveness. In *Proc. 18th Usenix Security Symposium (SEC'09)*. USENIX Association, 2009.

[62] Joshua Tan, Lujo Bauer, Joseph Bonneau, Lorrie Faith Cranor, Jeremy Thomas, and Blase Ur. Can Unicorns Help Users Compare Crypto Key Fingerprints? In *Proc. CHI Conference on Human Factors in Computing Systems (CHI'17)*. ACM, 2017.

[63] Anthony Vance, David Eargle, Jeffrey L. Jenkins, C. Brock Kirwan, and Bonnie Brinton Anderson. The Fog of Warnings: How Non-essential Notifications Blur with Security Warnings. In *Proc. 15th Symposium on Usable Privacy and Security (SOUPS'19)*. USENIX Association, 2019.

[64] WeAreDynamo.org. Guidelines for Academic Requesters. http://wiki.wearedynamo.org/index.php/Guidelines_for_Academic_Requesters, 2017.

[65] Tara Whalen and Kori M. Inkpen. Gathering Evidence: Use of Visual Security Cues in Web Browsers. In *Proc. Graphics Interface 2005 Conference (GI'05)*. Canadian Human-Computer Communications Society, 2005.

[66] WhatsApp LLC. WhatsApp - Features. https://www.whatsapp.com/features/, 2019.

[67] Alma Whitten and J. D. Tygar. Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0. In *Proc. 8th Usenix Security Symposium (SEC'99)*. USENIX Association, 1999.

[68] Justin Wu and Daniel Zappala. When is a Tree Really a Truck? Exploring Mental Models of Encryption. In *Proc. 14th Symposium on Usable Privacy and Security (SOUPS'18)*. USENIX Association, 2018.

[69] Yuxi Wu, Panya Gupta, Miranda Wei, Yasemin Acar, Sascha Fahl, and Blase Ur. Your Secrets Are Safe: How Browsers' Explanations Impact Misconceptions About Private Browsing Mode. In *Proc. 27th International Conference on World Wide Web (WWW'18)*. International World Wide Web Conferences Steering Committee, 2018.

[70] Philip Zimmermann. PGP Version 2.6.2 User's Guide. ftp://ftp.pgpi.org/pub/pgp/2.x/doc/pgpdoc1.txt, October 1994.

A Appendix

A.1 Demographics

Table 3 shows the demographics of participants in all studies.

	Study: Tool Usage	Study: Disclosures	Study: Icons	Study: Animations	Study: Validation
Participants					
Started	173	234	100	253	159
Finished	160	210	90	159	150
Valid ($n=$)	149	196	86	107	145
Gender					
Male	60.7%	54.1%	47.7%	67.3%	40.7%
Female	37.9%	44.9%	51.2%	29.0%	57.9%
Not M/F	1.4%	1.0%	1.2%	2.8%	1.4%
Ethnicity†					
White	78.5%	68.9%	68.6%	71.0%	89.0%
Asian or Pacific Islander	6.0%	14.3%	16.3%	7.5%	6.9%
Black or African American	13.4%	7.7%	11.6%	14.0%	0.0%
Hispanic or Latino	4.7%	12.2%	11.6%	10.3%	2.8%
Native American	0.7%	1.5%	0.0%	0.0%	0.0%
Other & Prefer not to say	0.7%	0.5%	1.2%	0.9%	4.1%
Smartphone OS†					
Android	67.1%	57.7%	41.9%	60.7%	59.3%
iOS	37.6%	50.5%	66.3%	39.3%	40.7%
Other	0.0%	1.0%	3.5%	0.0%	0.0%
No smartphone	1.3%	0.0%	0.0%	0.0%	0.0%
Prefer not to say	0.7%	0.0%	1.2%	0.0%	0.0%
Computer Science					
CS Education	28.9%	24.5%	24.4%	31.8%	26.9%
CS Job	22.1%	26.0%	29.1%	32.7%	19.3%
Age in years					
Mean	37.5	33.9	35.6	32.5	37.7
Std. dev. (σ)	10.7	9.0	10.9	8.4	10.5
Median	35.0	33.0	33.0	31.0	35.0

† Multiple answers allowed, may not sum to 100%

Table 3: Participant demographics.

A.2 Scale of Perceived Trust, Security and Privacy

Ten item scale of perceived trust, security and privacy. Participants choose from a 5-point likert scale on each question.

1. I think the new WhatsApp version is trustworthy.
2. I do not doubt the honesty of the new WhatsApp version.
3. I think the new WhatsApp version is secure.
4. I think only me and the recipient(s) can read our messages.
5. I think other people cannot send a message pretending to be me.
6. I think no one can unnoticeable modify messages sent between me and the recipient(s).
7. I think that if somebody hacks my phone, they will not be able to read my messages.
8. I think only me and the recipient(s) can know the messages were sent.
9. I think the new WhatsApp version does not collect more personal information than strictly needed.
10. I think the new WhatsApp version will not use my personal information for other purposes without my authorization.

A.3 Messenger Usage

The following figures show the messenger usage among our participants in the first survey. Figure 4 shows the preferred messenger for day-to-day conversations, Figure 5 shows the preferred messenger for sensitive or confidential conversations. Figure 6 shows all messengers used in the last 6 months.

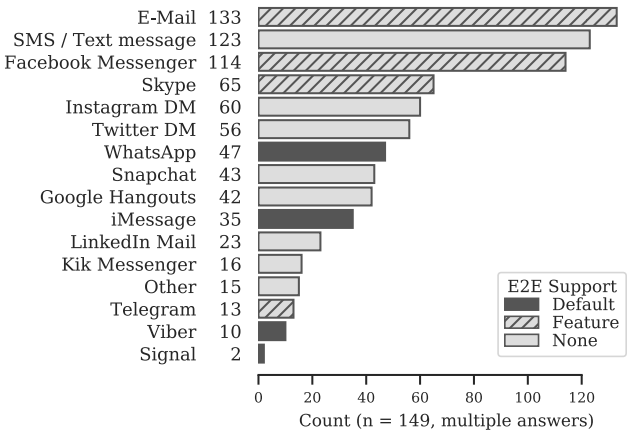


Figure 6: Study: Tool Usage - "Which online communication tools have you used in the last 6 months?"

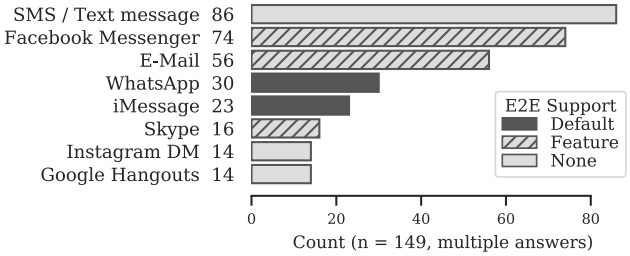


Figure 4: Study: Tool Usage - "Which tools do you prefer for day-to-day conversations?" (Top 8)

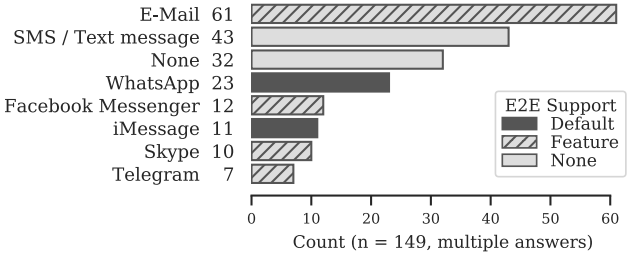


Figure 5: Study: Tool Usage - "Which tools do you prefer for sensitive or confidential conversations?" (Top 8)

A.4 Scripted Conversation

The text in Figure 7 was used in the scripted conversations that were shown to the participants in the videos. The overall screencast took on average 95 seconds.

A.5 Data Quality

Participant diversity and data quality on MTurk and Prolific is generally perceived as satisfactory [39–41]. We followed best practices [31, 41, 64] and required workers to be U.S. residents who have already completed 100+ HITs with a 95% approval rate.

During piloting, we experienced similar data quality issues as reported in recent work [31]. Therefore, we implemented a set of countermeasures, including blocking participants whose IP address came from outside the U.S. or belonged to a VPN or proxy service provider even within the U.S.[4]

Following best practices [30, 31], we added three attention checks to all questionnaires.

To remove a potential confound for Studies 2–5, we wanted to include only participants familiar with secure messaging apps. Therefore, we added an MTurk qualification task in which we asked participants which messaging apps they currently used and invited only WhatsApp users to the actual study itself. Workers who had participated in one study were ineligible for all subsequent ones. We paid each participant $0.15 for the short qualification task.

We took advantage of the pre-screening provided by Prolific, where participants had to report the regular use of WhatsApp in a pre-screening questionnaire[5], and required participants to be located in the U.S. or UK, have a 95% or higher approval rate and at least 100 previous submissions.

Me: Hi, darling. I forgot my wallet at home. Could you please look up my credit card number? I need to place an order before I forget.

<Wait 10s>

Remote: Sure, where is it?

Me: It should be on my desk.

<Wait 5s>

Remote: One second.

<Wait 25s>

Remote: It's 1234-5678-9012-3456, valid until 12/21, and the security code is 456.

Me: Thanks

Figure 7: Chat messages displayed in the application.

A.6 Study Video Introduction

We introduced the video to our participants in the following way:

[4] We used the https://iphub.info service to filter VPNs and proxies.

[5] The exact question we asked in the pre-screening questionnaire was: Which of the following chat apps do you use regularly? [multiple-choice]

Name (Alphabetical order)	E2E-Encrypted (Protocol Name)	E2E Indicator Color	Icon	Text	Platforms Android	iOS	Downloads (On Android)
E-Mail	○	-	-	-	●	●	-
E-Mail with PGP or S/MIME	● (PGP or S/MIME)	d	d	d	●	●	-
Facebook Messenger	◐ (Signal)	●	Lock	●[1]	●	●	5.000M+
FaceTime	● (SRTP)	○	○	○	○	●	-
Google Hangouts	○	-	-	-	●	●	5.000M+
iMessage	● (unknown)	●	○	○	○	●	-
Instagram DM	○	-	-	-	●	●	1.000M+
Kik Messenger	○	-	-	-	●	●	100M+
LinkedIn InMail	○	-	-	-	●	●	500M+
Signal	● (Signal)	○	Lock	○	●	●	50M+
Skype	◐ (Signal)	○	○	●[2]	●	●	1.000M+
SMS	○	-	-	-	●	●	-
Snapchat	○	-	-	-	●	●	1.000M+
Telegram	◐ (MTProto2.0)	●	Lock	●[3]	●	●	500M+
Twitter DM	○	-	-	-	●	●	1.000M+
Viber	● (unknown)	●	Shield[4]	●[5]	●	●	500M+
WhatsApp	● (Signal)	○	Lock	●[6]	●	●	5.000M+

● Yes (for E2E-Encrypted: Yes, by default) ○ No ◐ Has a "secret mode" which uses E2E encryption, but is not active by default [d] Depends on the client used [1] *Secret conversation* [2] *Private conversation* [3] *Secret chat* [4] *Has an additional Secret Chat, uses a lock icon* [5] *Messages sent in this conversation are encrypted* [6] *Messages to this chat and calls are now secured with end-to-end encryption*

Table 4: List of popular communication tools.

"*Imagine that WhatsApp will soon release a new version. This new version would have a different user interface in some places but have the same features as the version you are used to. Below we show you a brief video of what this new user interface for WhatsApp might look like.*"

A.7 Replication Material

The videos and questions used within this paper are available on our webpage at https://publications.teamusec.de/2021-soups-e2e/.

Concerned but Ineffective: User Perceptions, Methods, and Challenges when Sanitizing Old Devices for Disposal

Jason Ceci, Hassan Khan
School of Computer Science
University of Guelph
{jceci, hassan.khan}@uoguelph.ca

Urs Hengartner, Daniel Vogel
Cheriton School of Computer Science
University of Waterloo
{urs.hengartner, dvogel}@uwaterloo.ca

Abstract

Consumers are upgrading their devices more often due to continuous advances in hardware. Old devices need to be sanitized (i.e., personal data removed with low recovery probability) before selling, donating, throwing away, or recycling the device (*"disposal"*), but previous works have shown that users frequently fail to do that. We aim to understand the sources of misconceptions that result in risks to personal data. Through a survey (n=131), we measure where the old devices end up and how they are sanitized. Our survey shows that while most users dispose of their devices, a large proportion of participants (73%) kept at least one old device, often due to data leakage concerns. Among disposed-of devices, 25% of participants reported using methods to erase their data that are insecure. To further explore the processes that were undertaken to sanitize devices and sources of misconception, we invite a subset of respondents (n=35) for interviews. Our interviews uncover the reasons for poor device sanitizing practices—misleading data deletion interfaces and prompts, lack of knowledge, and complex and slow disk wiping procedures. We provide suggestions for device manufacturers and retailers on how to improve privacy, trust, and convenience when sanitizing old devices.

1 Introduction

The past decade has witnessed an explosive growth of personal electronic devices [11]. Most consumers own a smartphone and computer, and many have other devices, such as tablets, cameras, flash drives, and portable hard drives, which also store personal data [4]. With rapid advancements in technology, electronic devices have a relatively short life cycle, and users often upgrade these devices. For example, the average age of smartphones traded in is 3.2 years [21]. When upgrading, consumers have several options for what to do with their old devices, including selling, recycling, donating, discarding by throwing it away, or keeping it even without using it—we refer to these collectively as *"disposal methods"*. Current estimates show that more than 206 million used smartphones were sold worldwide in 2019, and this number is expected to grow to 332 million units by 2023 [14]. Recycling is another environmentally-friendly option to dispose of electronic devices and in 2019, 17.4% of old devices were recycled worldwide [6].

Disposing of old devices introduces privacy and data security risks since these devices often contain a variety of personal data including images, videos, messages, financial information, emails, and internet and location history. For this reason, prior to disposing of a device, the device should be "sanitized," which means removing the personal data on it with a low recovery probability [22]. Previous studies have shown that users often inadvertently leave some personal data or do not sanitize their devices at all before selling them. In a seminal work, Garfinkel and Shelat explored user data removal practices by examining 159 second-hand hard drives and found that 91% contained sensitive data [9].

More recent works have shown that this state of affairs has not changed much over the past two decades for smartphones or other personal devices [7, 23]. Despite the increasing availability of full-disk encryption on some types of devices, proper sanitizing is still important since the decryption key may be derived from a weak password only [10]. Therefore, it is critical to understand why users fail to sanitize their devices to stop this dangerous practice. To the best of our knowledge, our work is the first to explore device sanitizing practices entirely from users' perspectives.

We explore users' decision-making process starting from when they no longer need a device to discover any difficulties or misconceptions they may have in regards to sanitizing

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

their devices. For a holistic exploration, we consider different device types (smartphones, computers, tablets, cameras and drives), where each type differs in the data they contain and the ways to sanitize them properly. Some devices have simple sanitizing procedures (e.g., smartphones), while others may require multiple steps (e.g., cameras). Furthermore, devices may be encrypted, which may affect a user's perception of data privacy and sanitizing methods.

To capture these aspects, we conducted a two-part study. The first part was a survey (n=131) to establish what consumers do with devices when they no longer need them, as well as if and how users remove their data from their old devices. This investigates how people prepare devices for disposal across disposal methods, which has not been investigated previously. Next, in an attempt to understand why many users fail to sanitize their devices properly, we conducted semi-structured interviews (n=35) asking users about the steps they took to sanitize their devices to identify any difficulties or misconceptions. Users were also asked to rate how likely they believed data recovery was after they sanitized the device. This provides information on consumer confidence when they sanitize their devices and helps identify misconceptions. Our work is the first to use qualitative data to uncover users' perceptions, challenges they face, and misconceptions related to sanitizing devices. Our key findings include:

- The majority of respondents (61%) chose the "Factory Reset" feature on smartphones, computers, and tablets to sanitize the device. However, they had low confidence in the security of this feature, with 57% of respondents feeling it was extremely likely that an expert attacker could recover data from their devices. Consequently, 36% of users choose to keep old devices rather than sell or recycle them.
- Unsafe sanitizing practices were common—34% of interview participants reported manually deleting all or some of the data on their devices and considered it to be a secure disk sanitizing method. Manual deletion was also error-prone. During the interview, in 9/33 cases where manual deletion had been employed, participants admitted forgetting to sanitize certain types of data.
- We use our findings to revisit the plausible reasons for poor sanitizing practices for hard drives proposed by Garfinkel and Shelat [9]. Our evidence validates some reasons, including *lack of training, tool error*, and *hardware failure*. However, we found no evidence to support *lack of knowledge* or *lack of tools* as plausible reasons.
- Our interviews uncover other plausible reasons for poor sanitizing practices, including side effects of sanitizing and the time required for sanitizing.

Finally, we provide suggestions for device manufacturers on how and when to present the right information to users for more informed sanitizing decisions. We also highlight the importance of retailers sharing their device sanitizing practices for returned devices to better safeguard users' data.

2 Background and Scope

Before disposing of personal devices, confidential data can be removed in a variety of ways depending on the device type and operating system. Garfinkel and Shelat [9] define sanitizing as removing confidential information from storage before repurposing, retiring, or disposing of electronic devices. Furthermore, confidential information should be removed using processes that result in a low probability of recovery using existing data recovery tools and techniques.

Most modern devices (with the exception of cameras and drives) have a device sanitizing function often labelled as "Factory Reset," "Erase All Content and Settings," or "Secure Wipe" [24, 25, 27]. This function is generally designed to remove all user data and applications while leaving the operating system and factory-installed applications. Apple's support website for iOS explains that deleting files makes files inaccessible but does not remove them from the device, so before selling or giving away a device, "Erase All Content and Settings" should be run [25]. Apple's iOS Security Guide further explains that the "Erase All Content and Settings" option wipes the encryption keys to the user data, leaving all personal data inaccessible. On the other hand, Android and Samsung only mention that a "Factory Reset" will remove all data [24, 26]. Windows 10 provides two "Factory Reset" options and clearly explains what each does [27]. If the user chooses the "Data Erasure" option, it removes files and cleans the drive. It suggests when to use it ("If you're planning to donate, recycle, or sell your PC, use this option") along with a rough time estimate ("This might take an hour or two") along with the advantage ("... it makes it harder for other people to recover files you've removed").

However, these sanitizing functions differ between device types, operating systems, and underlying hardware, and the desired outcome depends on the device's encryption status. This is further complicated by the lack of public information on how a device is sanitized and whether the provided "Factory Reset" function properly sanitizes the device. Shu et al. [22] showed that the "Factory Reset" function on several Android devices fails to sanitize the device. Similarly, Wei et al. [28] identified challenges when sanitizing SSDs, including the requirement of invoking the SSD controller's secure erase function. They also demonstrated that the SSD controller's built-in secure-erase command often fails to sanitize the device.

Simply deleting files, like on a smartphone, is insecure as these files can be easily recovered. Deleting files removes only the record of that file in a table, leaving files data on the drive with the potential to be fully recovered [8]. Manually deleting data prior to disposing of the device is not recommended due to the high probability of data recovery. Standard methods for

formatting drives leave data on the drive recoverable using commercially available software [8].

Given the variety of devices and the lack of publicly available information, it is challenging to identify methods that properly sanitize devices. Therefore, in our study, we refrain from commenting on whether the "Factory Reset" or format option that users employed properly sanitized the device. We instead focus on users' perceptions, practices, and misconceptions when sanitizing old devices for disposal.

3 Related Work

Prior works have investigated specific areas of data deletion and consumer data privacy on second-hand devices and in the cloud. In this section, we discuss prior works that measure device sanitizing practices in-the-wild and users' perceptions of data deletion for online services. One line of research has explored approaches to securely delete digital data given different capabilities of adversaries, as well as how secure deletion approaches can be integrated into systems at different interfaces to protect against specific adversaries [20]. However, these works are only tangentially related since we focus on users' perceptions and practices when disposing of old personal devices. Interested readers are referred to Reardon et al. [19].

3.1 Device Sanitizing In-the-Wild

Previous research into personal data on disposed-of devices has focused primarily on measuring what proportion of disposed-of devices had recoverable personal data. The seminal work in this area was the 2003 study by Garfinkel and Shelat [9], which investigated personal data on second-hand hard drives purchased mostly through online auctions. They reported recovering large amounts of sensitive information. Only 9% of the hard drives had been properly sanitized (zero-filled) with most of the remaining drives having recoverable data, including 675 Microsoft Word documents and thousands of email messages and credit card numbers.

Several recent efforts have shown that the issue that Garfinkel and Shelat [9] identified has stayed the same. In a 2014 study, researchers from Avast procured 20 Android devices from eBay and found lots of sensitive data on them including pictures (with over 1000 pictures in various states of undress), contacts, chat logs, search history, and location history, among others. In a 2019 study, Jones et al. [15] purchased 100 second-hand phones via eBay and performed forensic analysis to show that 19% of the phones contained data from previous owners. This data included private emails, intimate photos, contact lists, text messages, tax documents, bank account details, web browsing histories, and personal calendars. The same team of researchers bought 100 used memory cards and showed that 67% of the cards had personal data of the previous owners on them [3]. Researchers from Ontrack and Blancco purchased and analyzed 159 used personal devices from the United States, Britain, Germany, and Finland to discover sensitive data on 42% of devices, with 15% containing personally identifiable information [29].

Most of these studies perform analysis of used devices to determine if there is recoverable personal data on them. Only Garfinkel and Shelat propose nine plausible user-centred explanations for the widespread data leakages on disposed-of devices. These include lack of knowledge, lack of tools, lack of training, tool error, and hardware failure (see Section 7). To the best of our knowledge, our work is the first to investigate these aspects entirely from the users' perspective.

3.2 Users' Perception of Data Deletion in Online Services

While no previous work has explored data sanitizing methods for disposed-of devices, researchers have investigated data deletion and expiration aspects from users' perspectives for online services. Ramokapane et al. [18] explored users' understanding and practices of deleting data from cloud storage or services. They found that the lack of information about deletion, incomplete mental models of the cloud and deletion within it, and poorly designed user interfaces for deletion functions lead to users' failure to delete data. Murillo et al. [17] conducted a study and two focus groups to understand online user data deletion, retention, and expiration. They found that the correct understanding depended on whether users think beyond the user interface or not. Habib et al. [13] investigated the usability and interaction paths of data deletion options for 150 websites. They found that while the majority of analyzed websites offered controls, they were inconsistent across websites and sometimes rendered unusable by missing or unhelpful information. Similarly, researchers have explored data deletion and expiration aspects for online social networks, including Twitter and Facebook [1, 2].

While these works may point out issues that may be true for secure data deletion in personal devices, such as missing or unhelpful information, consumers have a lot more control over data deletion processes on personal devices than cloud environments. Therefore, a focused effort needs to be carried out to investigate these aspects for personal devices.

4 Study Design

Our survey of related works shows that device sanitizing aspects from the users' perspective have largely been unexplored. Our main objective is to understand the privacy-related sanitizing practices when users dispose of their old devices. To achieve this objective, we explore the following questions:

- What do people do with their devices when they no longer need them?

- How do people perceive threats to sensitive data on the devices that they dispose of?
- What steps, if any, do people take to remove the sensitive data on their devices before disposing of them?
- Do people trust device sanitizing methods provided by device manufacturers?
- Are people aware of, and do they understand proper device sanitizing practices when disposing of their devices?

There are several challenges to this investigation due to permutations of devices, data, and possible sanitizing methods. First, different types of devices may store different types of data. For example, a smartphone may have personal pictures, whereas a laptop may only have work-related data. Second, different devices may store personal data differently. For example, a smartphone or laptop may store data in an encrypted format, unlike a digital camera with an external storage card. Third, different devices may support different ways to sanitize the device. For example, Windows 10 provides a "Secure Erase" feature, whereas a flash drive may require a tool to zero-fill.

To overcome these challenges, we conducted a two-part study. The first part consisted of a survey, which asked respondents about their device disposal and sanitizing practices. We captured this data across participants who rated themselves at different technology proficiency levels to capture misconceptions for novice, amateur, and technology-proficient users. This enabled us to obtain quantitative data on what consumers do with their devices when they no longer need them, as well as if and how they remove their data from their old devices. To obtain qualitative data, we conducted semi-structured interviews with a subset of survey respondents. We prioritized those participants who agreed to be contacted for the interview, disposed of devices across multiple device types, and represented a reasonable diversity of technology proficiency. The interviews enabled us to explore why many people fail to properly sanitize their devices before disposing of them.

In the following sections, we report the recruitment process, study procedure, and results separately for the online survey and semi-structured interview. Note that we received approval from our IRB for this study.

5 Online Survey

The goals of the survey are to understand how users dispose of their devices and what steps they take to sanitize their device. We limited electronic devices to smartphones, laptop and desktop computers, tablets, digital cameras, memory cards, hard drives, and flash drives as these devices are more likely to contain sensitive data.

5.1 Recruitment and Procedure

For the online survey, we recruited respondents by placing advertisements on Facebook marketplace, Kijiji (the Canadian equivalent of Craigslist), local subreddits, and through word-of-mouth (see advertisement text in Appendix A.1). The inclusion criteria were that the respondents should have recently listed an item for sale on an online buying and selling marketplace. Participants responded to the survey on Qualtrics (see Appendix A.2). The survey collected data from respondents for the following data categories: (a) Demographics and background; (b) Disposal methods and reasons for not disposing of; (c) Sanitizing methods; and (d) Sanitizing perceptions.

At the end of the survey, respondents were asked if they wish to be contacted for a follow-up interview. For the 10-minute online survey, participants were paid $2.

Table 1: Survey Demographics (*UD = Undisclosed)

n = 131							
Gender							
Woman	*Man*	*Other*	*UD*				
68	57	2	4				
Age (in years)							
18–25	*26–30*	*31–35*	*36–40*	*41–45*	*46–50*	*50+*	*UD*
38	27	20	14	15	7	6	4
Self Reported Proficiency in Technology							
Basic	*Intermediate*	*Advanced*	*UD*				
8	86	34	3				

5.2 Results

For test statistics on quantitative data, Pearson's Chi-Squared test was used to compare categorical data, a Kruskal-Wallis one-way analysis of variance was used to compare Likert scale responses between respondent technology proficiency levels, and Wilcoxon Signed-Ranks test to compare Likert scale responses between questions.

5.2.1 Demographics and Background

Respondents were asked about their age, gender, and their level of technology proficiency. The survey was completed by 131 participants. Their demographics are summarized in Table 1. It shows reasonable diversity among respondents in terms of gender. In terms of age, while almost half of the respondents (65/131) are 30 years of age or younger, we have a good representation from other age groups as well. Fewer respondents self-reported a basic proficiency in technology as compared to those who self-reported as intermediate or advanced. This smaller proportion is somewhat expected due to the study being advertised and conducted online and respondents possibly over-reporting their technology proficiency.

5.2.2 Disposal Methods

We asked respondents what they did with electronic devices they no longer used including smartphones, computers, tablets,

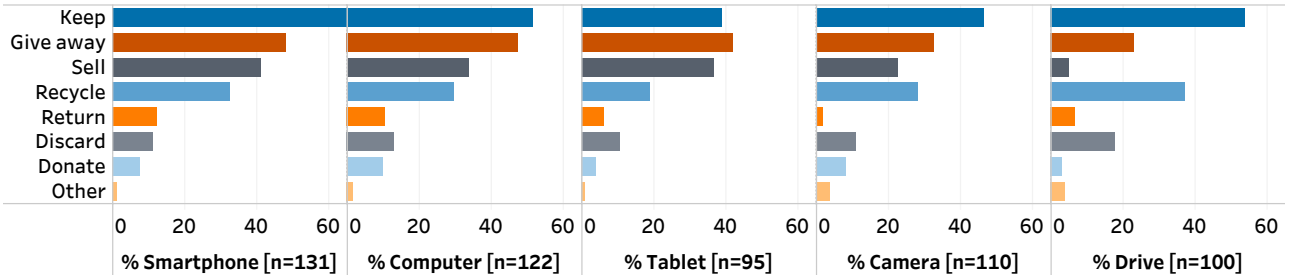


Figure 1: Response of participant ('n') to "What do you do with electronic devices you no longer use? (Choose all that apply)"

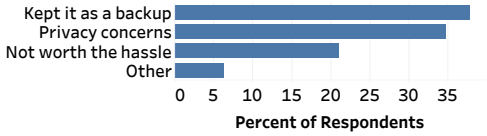


Figure 2: Respondents' responses to "What was the main reason you kept an old electronic device?"

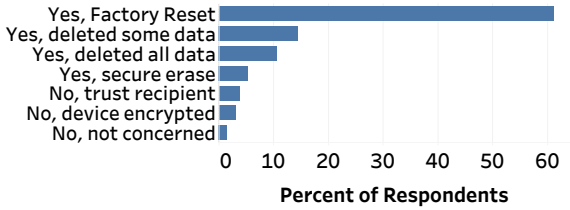


Figure 3: Respondents' responses to "Did you remove any personal data on your old device before selling, giving away, recycling, donating, or returning it?"

digital cameras, hard drives and flash drives (see Figure 1). Keeping the device was the most reported action for all device types (except for tablets), with 73% of respondents keeping at least one device. Giving the device to a friend or family member was the second most reported way to dispose of smartphones, computers, and cameras/memory cards. Giving the device to a friend or family member was the most reported action for tablets and the third-most reported for drives (we use the term "drives" to refer to hard drives and flash drives). Selling and recycling old devices was a common action for all device types, and respondents reported selling more smartphones, computers and tablets than recycling. For cameras and drives, respondents reported recycling more than selling. For each device type, at least ten respondents reported throwing out a device. While some respondents reported returning their smartphones and computers to a provider or IT department, understandably, this action was rare for cameras and drives. The "other" responses were codified, and all responses referred to some way of destroying the device or its storage medium.

Of the 95 respondents who reported keeping an old electronic device, 38% (36/95) reported keeping it as a backup, 35% (33/95) kept it due to privacy concerns, and 21% (20/95) kept it mainly because it was not worth the hassle to sell, donate or recycle (see Figure 2).

5.2.3 Sanitizing Methods

Respondents were asked if they removed their personal data from the last device they sold, donated, recycled, or returned. We only focused on the last device to simplify the survey and maintain reasonable time constraints. If they did attempt to remove any personal data from the device, they were asked to select the method they used and whether all or only some personal data was removed (see Figure 3). 25% (33/131) used a method that is known to be insecure, such as manually deleting some or all data on the last device they sold, donated, recycled, or returned. Few respondents (11/131) chose not to try to remove their personal data. 62% (80/131) of respondents reported that they used a built-in "Factory Reset" function. Seven respondents used a tool or utility to zero-fill or secure erase the data storage. A Chi-squared test found no significant effect for technical proficiency (basic, intermediate and advanced) and whether they chose to use a secure ("Factory Reset", zero-fill or secure erase) sanitizing method or chose to manually delete data ($\chi^2(2) = 1.00, p = 0.61$).

5.2.4 Sanitizing Perceptions

Respondents were asked how concerned they would be if an untrusted individual was able to access their data on an old device on a 5-point Likert scale. Most respondents reported that they were "most concerned" (60% (78/131)) or "concerned" (24% (32/131)) about their data being accessed while (13%) 17/131 were "somewhat concerned", (2%) 3/131 were slightly concerned, and only (1%) 1/131 was least concerned. A Kruskal-Wallis test examined the effect of respondent technical proficiency on the reported level of concern and found no significant differences ($H(2) = 1.33, p = 0.51$).

Respondents were asked how likely they felt it would be for two theoretical attackers, one with average and one with expert computer skills, to recover any data from their device after their chosen sanitizing method. On a 7-point Likert scale,

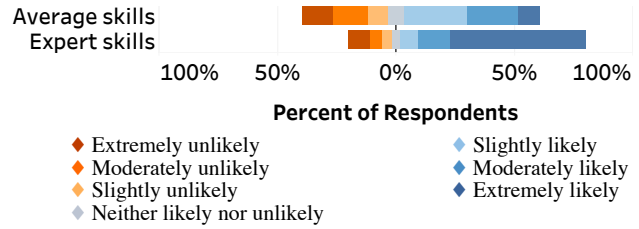


Figure 4: Respondents' responses to "How likely do you believe it would be for a person to be able to recover any personal data from the last device you sold, donated, recycled or returned?"

(57%) 75/131 of respondents felt it was at least slightly likely that an attacker with average computer skills could recover their personal data (see Figure 4). With an expert attacker, (57%) 75/131 respondents felt it was extremely likely that the attacker could recover their personal data. A Wilcoxon Signed-Ranks test indicated that likelihood ratings for an expert attacker were significantly higher than the likelihood ratings for an average attacker ($Z = 8.31, p < 0.001$).

6 Semi-Structured Interview

The goal of the semi-structured interview was to explore why users fail to sanitize their devices properly. All interview participants had already completed the survey, and the interview was treated as an extension to the survey.

6.1 Participants

Respondents who expressed their interest in participating in the follow-up interview and met the inclusion criteria (i.e., had disposed a device across two or more device categories and were interested in a follow-up) were invited to participate. 66 survey respondents met the inclusion criteria and were contacted over email to participate in the interviews (of which two declined, and 29 did not respond).

6.2 Procedure

Due to the pandemic, the interviews were conducted online (using Google Hangouts or Skype). We chose a platform that supported video and screen sharing as it allowed us to share screenshots of common device sanitizing interfaces (more details to follow). We performed audio recording if the participant consented. Otherwise, the researcher took notes. For participating in the interview, participants were paid $20. The interview questions were broadly categorized into the following categories and required both categorical and free form responses (also see questions in Appendix A.3):

- **Demographic and background:** We sought further demographic information and general questions about their electronic devices and the data stored on them.
- **Sanitizing methods:** We asked participants if and how they removed their personal data before the following (when applicable): giving away the device to a friend or family member, selling the device, donating the device, recycling the device and returning the device to a provider, manufacturer, workplace or IT department.
- **Sanitizing non-functioning devices:** We asked participants what they do with the devices that they no longer use that are broken or defective. If they sold, donated, recycled, or threw away the device, they were asked if they attempted to remove any personal data and how.
- **Finding data:** Participants were asked if they ever found another person's data on a device they had purchased.
- **Challenges and misconceptions in device sanitizing:** We asked participants how difficult they think it is to fully remove all data from the different device types included in the study. We also explored the information presented to them when they were using "Factory Reset" or wipe procedures offered to them by the device manufacturer. To refresh their memory, we showed them possible user interfaces (specific to their device/OS) that are presented to the user when they are factory resetting their device. We collected these user interfaces for all common platforms, and they included Disk Format across common platforms (Windows, Mac, Linux) and "Factory Reset" interfaces for Windows, iOS, and Android.
- **Responsibilities:** Finally, we asked participants about the level of responsibility they feel online marketplaces, device manufacturers, and consumers themselves should have when it comes to practices around device sanitizing.

Table 2: Interview Participant Demographics (*UD = Undisclosed)

n = 35							
Gender							
Woman		*Man*		*Other*		*UD*	
20		14		0		1	
Age (in years)							
18–25	*26–30*	*31–35*	*36–40*	*41–45*	*46–50*	*50+*	*UD*
10	9	4	4	3	4	1	0
Annual Household Income (× $1000)							
<$30		*$30–74*		*$75–99*		*>$100*	*UD**
5		14		6		8	2
Highest Education Level							
High School		*Undergraduate*		*Graduate*			
10		20		5			
Self-Reported Proficiency in Technology							
Basic		*Intermediate*		*Advanced*			
4		29		2			

6.3 Results

For qualitative analysis, two researchers independently performed open coding to identify codes or themes in participant responses to free-response questions (Q9, 14, 16, 20, 22 in Appendix A.3). Identified codes were compared and discussed by reviewers until consensus was reached. For the qualitative data from interviews, we report quotes from participants when they represent a theme. In this case, we identify the number of participants who expressed that theme and provide a representative quote. When reporting quotes from participants, the participant number corresponds to the respondent number in the survey.

Table 2 summarizes the demographic information for participants of the semi-structured interviews. Additionally, annual household income levels and highest education level achieved are reported for the interview participants. Overall, our participant pool has good diversity for these demographics, which is important for our investigation.

6.3.1 Sanitizing Methods when Giving Away

Participants were asked if and how they sanitized devices that they gave to a friend or family member. 60% (12/21) reported using a built-in "Factory Reset" option citing its ease, speed, security, and availability. 24% (5/21) manually deleted all personal data and stated that they deleted their contacts, text messages and photos. When asked about certain types of data, 3/5 indicated that they forgot to remove their browsing history, saved passwords, or saved passwords.

> *"I just deleted everything I could find on the phone like apps, contacts, photos and videos. I didn't think about any website passwords or browsing history."* (P89)

For the giving away computers case, 47% (7/15) reported manually deleting data. 5/7 chose this method because it was the only method they knew. 2/7 had the knowledge of more secure methods but did not employ them because they trusted the recipient. Due to trust, 2/15 and 2/15 participants reported only deleting some personal data and no data, respectively. 20% (3/15) participants used the Windows "Reset My PC" function. The last participant manually deleted personal data before wiping the free space with a commercial tool.

When giving away tablets, one participant did a "Factory Reset" while one did not remove any data due to trust in the recipient. For digital cameras, 2/4 formatted the memory card using the camera's user interface, 1/4 used the "Select All" and delete option in the camera, and 1/4 did not remove anything due to trust in the recipient. For drives, 1/4 participants manually deleted data, 1/4 participants formatted the drive, and 2/4 participants did not remove any data due to trust.

6.3.2 Sanitizing Methods when Selling

We asked participants how they sanitized devices prior to selling them (see Figure 5). For smartphones, 93% (14/15) reported using "Factory Reset". However, 53% (8/15) reported manually deleting all personal data before selling computers. 5/15 who reported using a manual delete method did so as they did not know of a better method.

> *"I backed up files and then deleted them to the recycle bin. I didn't know there was anything else to it. I couldn't take the hard drive out because then the laptop won't work."* (P63)

3/15 computer owners considered more secure sanitizing methods but ultimately chose to delete their data as more secure methods were too technical or too slow.

> [On selling their Windows 7 laptop] *"I just deleted all my documents and photos. I couldn't find a factory reset button, so I googled it but wiping it was complicated to do."* (P36)

Only four participants reported selling their tablets —two used a "Factory Reset" option, one deleted some personal data manually, and one participant did not delete anything as they were not concerned about their personal data. With digital cameras, two participants deleted all photos manually, and two participants sold the camera without a memory card. No participants reported selling drives.

Participants were asked on a 7-point Likert scale how likely they felt it would be for two theoretical attackers, one with average and one with expert computer skills, to recover any data from the different device types that they sold (see Figure 6a and Figure 6b, respectively). Note that, to avoid priming the participants, we chose not to define the "average" and "expert" attackers. While participants may attribute impossible abilities to the expert attacker, this question was intended to gauge the confidence of participants in the security of their chosen sanitizing method. For smartphones, when participants were asked about an average attacker, 50% (8/16) participants reported a successful attack to be slightly or extremely unlikely, whereas the remaining participants reported it to be at least slightly likely. For an attacker with expert computer skills, 37% (6/16) participants felt it was "extremely likely" that they could recover data.

More than half of the participants felt data recovery was likely on computers by both types of attackers (75% (12/16) perceived likelihood of an expert attacker recovering data was at least slightly likely). For cameras, the likelihood of data recovery was reported as "extremely unlikely" or "moderately unlikely."

6.3.3 Sanitizing Methods when Recycling or Donating

Participants were asked if and how they sanitized devices that they had donated or recycled in the past. Participants reported donating 23 devices in total and at least one device for each device type. Participants reported using a "Factory Reset" method on 58% (7/12) smartphones, 1/5 computers, and 1/2 tablets before disposal. 2/5 participants reported removing the hard drive of a computer before donating, and one participant

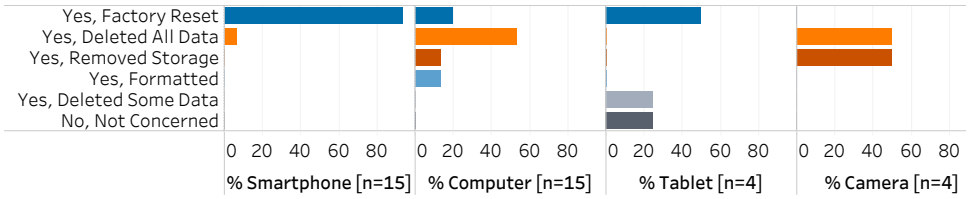
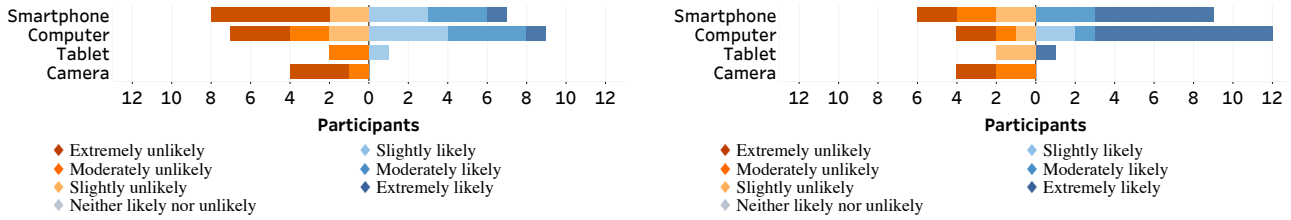


Figure 5: Participants' responses to "Did you remove any data from the device before selling it?"



(a) Average attacker.

(b) Expert attacker.

Figure 6: Participants' responses to "How likely do you believe it would be for a person with (a) average or (b) expert computer skills to be able to recover any personal data from the last device you sold?"

reported reinstalling the OS of the tablet before donating. The rest of the participants either reported using unsafe methods or were not concerned about threats to personal data. For a detailed breakdown, see Figure 9 in Appendix A.4.

Unlike when selling, giving away or returning a device, the device does not have to be functional when donating or recycling the device. Four participants, two with smartphones and two with digital cameras, reported not removing any data from their devices prior to donating or recycling them.

> *"The battery was dead, and I didn't have the charger, so I just donated it as is"* (P15)

We further explore users' sanitizing behaviours with non-functioning devices in Section 6.3.5.

6.3.4 Sanitizing Methods when Returning

We explored sanitizing practices for devices that are returned to a service provider, IT department, manufacturer, or retailer. For smartphones, 3/10 participants reported using a "Factory Reset" while 1/10 reported wiping the device in the recovery mode. 3/10 reported manually deleting some or all data. 3/10 also reported not removing any data because either they felt that it did not have personal data on it or they believed the store would sanitize the device.

> *"I returned my iPhone to the store to upgrade to a new model, like a trade-in. They said they wipe them, so I just gave it to them without doing anything."* (P67)

Seven participants returned computers to IT departments or retailers. 5/7 reported not removing any data because removal was too inconvenient, and they hoped that the retailer would sanitize the device before reselling them.

> *[On returning their laptop to the retailer] "I was going to remove some data, but it got way too inconvenient to try and delete everything before returning it for a trade-in. I think they delete everything there."* (P90)

2/7 reported removing all their personal data manually. When asked about how they manually removed all their personal data, one participant responded:

> *"I made a new user account and removed the old one in the control panel. I had to leave the laptop after an internship, but I had put a lot of my personal stuff on it, and it was linked to my phone."* (P21)

This method of removing personal data was a unique one in this study. For a detailed breakdown, see Figure 10 in Appendix A.4.

6.3.5 Sanitizing Methods for Non-Functioning Devices

Participants were asked how they disposed of and sanitized non-functional devices. We defined non-functional devices as devices that were broken, damaged, or defective in a way that impacted the usage of the device, including damaged screens, defects in data storage, battery or input, and devices that would not power on. Participants' responses are summarized in Figure 7. Across all device types, the majority of the participants chose to keep their non-functional devices. Less popular choices included throwing away the device, selling the device or destroying the device. Of participants that chose to keep non-functioning devices, 60% kept them out of concern for their personal data.

> *"I kept my broken tablet because I had my kid's pictures on it. I kept it for privacy reasons."* (P30)

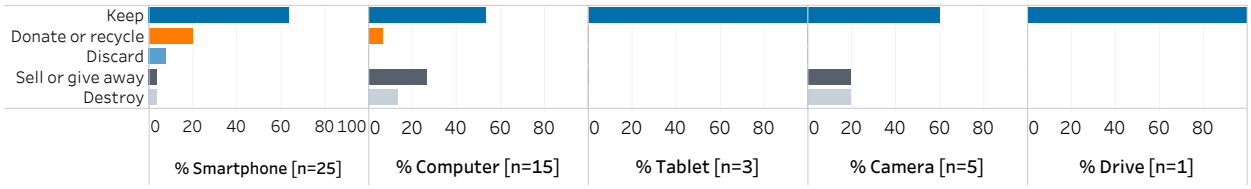


Figure 7: Participants' responses to "What do you do with non-functioning (broken) devices that you no longer use?"

For some participants, keeping non-functioning devices became impractical due to the space used, resulting in the devices being recycled without any sanitizing and with the potential for a personal data breach.

> *"My wife wanted me to get rid of our pile of old smartphones, so I just threw them all out. They wouldn't turn on anyway; the batteries were dead. I have no idea what was left on them; some are my daughter's and son's phones, so I don't know."* (P129)

Participants that had disposed of a non-functioning device were asked if they attempted to sanitize the device first. For smartphones, 4/9 participants attempted to remove data, while only 2/7 participants attempted to remove data from their broken smartphone prior to disposing of it. For a detailed breakdown, see Figure 11 in Appendix A.4.

6.3.6 Data Left on Resold Devices

We ask participants if they had ever bought a device that had the personal data of the previous owner on it. 12 participants (34%) reported finding the previous owner's personal data on a device that they purchased. Personal data that was reported to be found included photos, documents, and application login credentials. These findings validate previous reports of a major retailer selling a refurbished laptop with the previous owner's personal data [5]. 4/12 participants reported purchasing these devices from major electronics retailers.

> *"I had bought an open box laptop from [a major retailer] that had a lot of someone's files like photos and documents. Their OneDrive account was also logged in on the laptop."* (P67)

6.3.7 Participant Views on Device Sanitizing

To further understand participants' perception of the device sanitizing process, we asked them the level of difficulty ("Easy", "Intermediate", "Difficult/Impossible") they faced when they sanitized a device in the past for all the device types that they sanitized in the past (see Figure 8). For smartphones, 43% (15/35) participants felt it was "Easy", 46% (16/35) felt it was "Intermediate" and 11% (4/35) felt it was "Difficult/Impossible" to securely remove all data. For tablets,

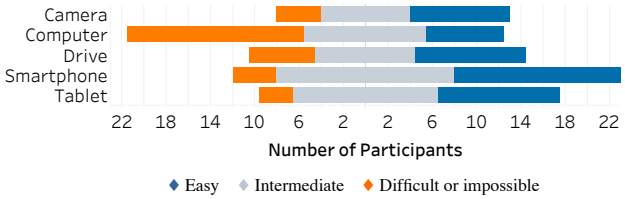


Figure 8: Participants' responses to "How difficult do you believe it is to securely remove all data from these devices?"

participants' responses were similar to those for smartphones. For computers, only 20% (7/35) found it "Easy," 31% (11/35) found it "Intermediate," and 49% (17/35) found it "Difficult/Impossible. For digital cameras and drives, 43% (9/21) and 40% (10/25) participants felt it was easy to securely remove all personal data, respectively. However, no participant previously reported using a secure sanitizing method (excluding participants that removed the memory card). This indicates that secure sanitizing procedures are not widely known for digital cameras and drives, and as a result, users falsely believe that manually deleting all data is adequate for sanitizing.

To assess the impact of usable device sanitizing methods, we asked participants if the difficulty to securely remove data has ever made them reluctant to sell or donate a device. 74% (26/35) participants answered "yes," and 26% (9/35) answered "no." 57% (20/35) strongly agreed that they would be more likely to purchase a device that they knew had a feature to securely remove all personal data, compared to one that did not. Finally, to determine the impact of a secure data removal feature on the future sale of a used device, participants were asked if they would be more likely to sell a device if it had a feature to securely remove all personal data. 77% (27/35) strongly agreed, and 14% (5/35) somewhat agreed.

Participants were asked whether the information conveyed by the "Factory Reset" feature assisted them when they were selling their device. Only 7/30 and 3/23 participants agreed that it assisted them when selling their smartphones or computers, respectively. For a detailed breakdown, see Figure 12 in Appendix A.4. Participants were asked to determine where the responsibility lay for their data privacy when selling devices. 94% (33/35) somewhat or strongly agreed that online marketplaces should explain the risks associated with selling used devices and how to securely sanitize used devices. 80%

(28/35) strongly believed that device manufacturers should make it easier to securely remove all personal data.

7 Discussion

We summarize and apply our results in two ways. First, we revisit the reasons for poor sanitizing practices among users hypothesized by Garfinkel and Shelat [9] in light of our findings. Second, we report new reasons for poor device sanitizing. Finally, we provide suggestions for improvement.

7.1 Revisiting Garfinkel and Shelat

Garfinkel and Shelat propose nine possible reasons why users frequently fail to sanitize their disk drives. While we cover devices of different types, their possible reasons may be applicable in a broader sense given the nature of storage mediums. We use the qualitative and quantitative findings from our study to validate their explanations (quoted verbatim in headings and italics).

Lack of knowledge. *("The individual simply fails to consider the problem.")* We find no evidence to support lack of knowledge as a source of poor device sanitizing practices. While participants displayed a *lack of concern* and *lack of training or incompetence*, all participants appeared to understand the problem well. It is plausible that awareness of the problem has increased due to media coverage or the overall improvement in the technology proficiency of the population.

Lack of concern for the problem. *("The individual considers the problem, but does not think the device actually contains confidential information.")* We find some evidence for this: 2% (3/131) survey respondents reported not removing data from the last device that they disposed of (to untrusted recipients) because they were not concerned. Interview participants were asked this question for each device type, and nine participants reported disposing of a device without removing data for at least one device type because they were not concerned. The codified responses indicate that 3/9 participants did so because they did not consider data to be sensitive, for example:

> *"I bought a new laptop but it broke and I had to return it for a new one. I only had it for a few weeks so I wasn't concerned"* (P60)

Lack of concern for the data:. *("The individual is aware of the problem—that the drive might contain confidential information—but doesn't care if the data is revealed.")* During the survey, no participant answered "least concerned" if an untrusted individual was able to retrieve data off of their devices and only 3% answered "slightly concerned." The rest (97%) answered "moderately" to "very concerned." However, the interview responses of participants who disposed of a device because they were not concerned, suggest otherwise. 3/9 participants indicated that they did so because of the lack of concern for data. Their responses were similar to:

> *"I was not concerned with the info on this phone, I just used it for calls, emails and texts, what would anyone do with that? "* (P129)

Failure to properly estimate the risk. *("The individual is aware of the problem, but doesn't believe that the device's future owner will reveal the information")* 3/9 interview participants who did not sanitize their device reported not being concerned partially due to the hope that the future owner will sanitize it and not reveal their information. Their responses were similar to:

> *"I only used it for web browsing and email so I didn't care about erasing it. I think the person who bought it will reset it."* (P23)

Seven participants also reported that they relied on retailers to sanitize their returned devices. This choice is also influenced by other factors, such as the amount of effort required to sanitize. Their comments were similar to:

> *"I was going to remove some data but it got way too inconvenient to try and delete everything before returning it for a trade-in. I think they delete everything there."* (P90)

Despair. *("The individual is aware of the problem, but doesn't think it can be solved.")* The majority of survey participants (60%) believed that a person with average computer skills would be able to retrieve data from a sanitized device. Despite the belief that sanitizing methods were imperfect, participants still sanitized their devices. During the interview, we asked participants how difficult they believe it was to fully remove all personal data from devices ("Easy," "Intermediate," "Difficult/Impossible"). While 4/35, 4/35, and 3/35 participants reported it to be "Difficult/Impossible" for smartphones, cameras, and drives, respectively, 20 participants reported it to be "Difficult/Impossible" for computers. This despair may be driven by a lack of training.

Lack of tools. *("The individual is aware of the problem, but doesn't have the tools to properly sanitize the device.")* Since Garfinkel and Shelat's work, free disk sanitizing tools have become widely available. However, these tools may not be readily accessible to the users and resources and information about these tools may be difficult to comprehend by non-expert users. These aspects are more related to the training or competence of the users and are discussed below.

Lack of training or incompetence. *("The individual attempts to sanitize the device, but the attempts are ineffectual.")* Even though deleting data does not sanitize disks, 25% of survey respondents deleted some or all data from their devices before disposal. 34% of interview participants made the same mistake and trusted this unsafe method for device sanitizing. During interviews, five participants reported that deleting files was the only sanitizing method known to them. Even with the knowledge that manual data deletion may not be the best option, one participant reported using it due to their inability to make a more secure method work:

[On selling their Windows 7 laptop] *"I just deleted all my documents and photos. I couldn't find a factory reset button so I Googled it but wiping it was complicated to do"* (P36)

Tool error. *("The individual uses a tool, but it doesn't behave as advertised.")* The wide use of manual delete coupled with the misunderstanding of participants that it is a secure device sanitizing choice is due to tool error. Four participants even reported using manual delete specifically because they believed it was secure. The prompt shown to users when emptying the "recycle bin" on Microsoft Windows reads: "Are you sure you want to *permanently* delete all of these items?"(emphasis ours). This seemed to be a source of confusion for several participants.

"It said it would be permanently deleted if I emptied the recycle bin" (P104)

Garfinkel and Shelat make this observation as well, noting that users falsely believed the format command removed all data from the drive because of the warning that "ALL DATA ON NON-REMOVABLE DISK DRIVE C: WILL BE LOST".

Hardware failure. *("...a computer failure might make it seem that the hard drive has also failed, when in fact it has not")* Six interview participants reported disposing of devices that no longer function without attempting to sanitizing them. However, these participants acknowledged that the device may still contain data, but believed it would be unlikely for someone to recover the data on a broken device with comments like:

"I had a laptop with a broken trackpad. I recycled it at [Major Retailer]. It was too hard for me to delete anything with the keyboard and the battery was dead so I just recycled it without the charger hoping it wouldn't be recovered." (P68)

7.2 Barriers to Secure Sanitizing

Our study shows that the majority of Garfinkel and Shelat's plausible explanations contribute to poor device sanitizing practices. We also note the following contributing factors.

Side effects of sanitizing. During interviews, 2/35 participants complained that if they were to use an existing tool to secure erase their computer, they would be left with an unbootable computer that would be difficult to sell. Reinstalling the operating system and drivers is outside the expertise of most users. Additionally, most computers now ship without the operating system installation disks and instead rely on recovery partitions on the computer's hard drive. Using disk wiping software would effectively erase this partition requiring the creation of installation media on another computer (which the user may not have access to) [12]. For example:

"There is a wipe software called DBAN my IT friend said to use but if I use that the computer won't boot anymore because Windows will be wiped out. I wouldn't have sold it if it didn't work."(P19)

Removing the storage media before disposing of the device is a secure method. However, three participants reported issues selling or donating without the storage media, such as the device no longer being operational or having less resale value. One participant commented:

"I used the delete all button on a Canon camera before donating. I donated it with my memory card so someone could actually use it."(P109)

Slow sanitizing process. Secure erase requires zero-filling storage media (the process for SSDs is more complicated and involved (see Wei et al. [28]). Five interview participants choose to use a less secure sanitizing method because a more secure one would take too long. For example:

"It takes way too long to delete everything, even removing programs took forever so I deleted the "My Documents" folder then gave it away." (P68)

Missed data. While manually deleting data is not a secure erase method, it provides some protection against new owners who are not actively looking for data remnants. During the interviews, participants who manually deleted data were asked about their deletion process and how they deleted certain types of data. Their responses indicate that while they deleted their personal data, they often forgot about data saved in applications such as browsing history, saved passwords in the browser's password manager, and application credentials. This oversight was reported by seven participants.

7.3 Improving Device Sanitizing Practices

When discussing potential improvements, we focus on widely used device types and platforms, specifically: Windows and macOS for computers and Android and iOS for smartphones. We discuss potential ways device manufacturers, retailers, and used marketplaces can help improve sanitizing practices.

Device Manufacturers. Device manufacturers can influence poor sanitizing practices that are due to the lack of training or incompetence and tool errors. When emptying the "recycle bin" or formatting the drive on Windows 10, users are informed that the "data will be permanently deleted." iOS (v14.4) and Android (v10) provide similar messages ("This photo will be deleted. This action cannot be undone." and "Permanently delete 1 image?", respectively). These messages warn users so they do not accidentally delete data making it (possibly) non-recoverable. However, such prompts create confusion from a sanitizing perspective. Apple's support website provides a more informed message saying: "...deleting files makes files inaccessible but does not remove them from the device" [25]. A message that warns users about their data being inaccessible without creating confusion regarding

device sanitizing is much needed. However, more exploration is needed to ensure that such a message is appropriate for users with different technology proficiency levels.

The message provided on Apple's support website is informative, but it is not presented to users at the *right* moment. This message needs to be presented to users when they empty the "recycle bin" or format devices or operating systems. Perhaps AI tools could be used to address this gap by detecting patterns of deletion that characterize device sanitizing and suggesting secure erase methods. Another possibility is to remind users to sanitize their devices when they unlink accounts from a device, as this is often done before disposal.

Finally, the information provided for the "Factory Reset" method needs reconsideration. During interviews, only 7/30 and 3/23 participants agreed that the information provided by the "Factory Reset" method assisted them when selling their smartphones or computers, respectively. The support websites of Android and Samsung only mention that a "Factory Reset" will remove all data without providing further details [24, 26]. Apple's iOS support page provides more details about the secure erase of the decryption key. The "Reset Device" interface of Windows 10 provides an elegant solution. The user is presented with two options—"Data Erasure on" or "Data Erasure off". Between them, it clearly explains the purpose (device reset due to issues vs. preparing to sell) and advantages (fast vs. making data recovery difficult) of each. Such options may help users make more informed choices.

Retailers and Used Marketplaces. During the interviews, several users said that they relied on the retailers to sanitize their used devices after they returned or exchanged them. However, the policies and procedures adopted by retailers to sanitize the device are not communicated to people and are imperfect [5]. When accepting used devices, retailers should provide information regarding how devices will be sanitized, potentially informing about the data erasure standard that will be followed (e.g., NIST 800-88 [16]).

Participants were asked to determine where the responsibility lay for their data privacy when selling devices. 94% (33/35) somewhat or strongly agreed that online marketplaces should explain the risks associated with selling used devices and how to sanitize used devices. While arguably these marketplaces have an ethical responsibility to inform the sellers about potential risks and ways to sanitize their devices, the economics of this action needs more investigation. On the one hand, transparency about such risks may stop sellers from selling their devices and on the other hand, with the availability of information on how to sanitize the devices, more people may be willing to sell their used devices.

8 Limitations

This study has several reasonable limitations intrinsic to studies on human participants. First, it relies heavily on self-reported information, which may be limited by the participants' memory or subjective views. Second, participants may be inclined to provide biased responses to avoid embarrassment or to provide what they feel are favourable responses. Furthermore, due to the pandemic, advertising for and participation in the study was exclusively online. This may have reduced the number of older participants or participants with basic technology proficiency. As these limitations are not easily preventable, we focus on limitations specific to this study.

During the study, there was a gap of up to a month between the survey and the interview for some participants. This gap was introduced since we waited for all surveys to be completed before the interview. During this gap, participants may have changed their device sanitizing methods. However, as the majority of the devices discussed in the interviews were sold before the survey, this gap is unlikely to have a significant impact on our findings. Furthermore, participants reported their disposal methods and experiences that happened several months ago. Their recall of the disposal methods may not have been accurate. While we presented them with "factory reset" interfaces for specific platforms (if needed), they likely experienced an interface with some variation. Finally, the behaviour of some participants may be influenced by the presence of device encryption, but we did not explore this. Exploring how device encryption changes users' behaviour is a possible future work.

9 Conclusion

We conducted a survey with 131 participants and a semi-structured interview to understand why users adopt unsafe practices when disposing of their old devices. Our investigation provides evidence that the unsafe practices are due to the lack of knowledge, misleading prompts and descriptions provided by device manufacturers, time constraints, possible side effects of sanitizing, and the delegation of sanitizing to retailers without a clearly defined policy from them. Our study provides little or no evidence for some of the previously suggested reasons for improper device sanitizing practices. Finally, we suggest possible improvements in user prompts and descriptions that can be adopted by the device manufacturers to mitigate the misconceptions of users. With the more frequent disposal of devices containing personal data, our findings will help researchers, device manufacturers, and retailers improve device sanitizing practices for consumers.

Acknowledgements

This material is based upon work supported by NSERC under Grant No. RGPIN-2019-05120. We thank Jonah Stegman for his feedback on the survey and assistance.

References

[1] Hazim Almuhimedi, Shomir Wilson, Bin Liu, Norman Sadeh, and Alessandro Acquisti. Tweets are forever: A large-scale quantitative analysis of deleted tweets. In *2013 Conference on Computer Supported Cooperative Work*, pages 897–908, 2013.

[2] Lujo Bauer, Lorrie Faith Cranor, Saranga Komanduri, Michelle L Mazurek, Michael K Reiter, Manya Sleeper, and Blase Ur. The post anachronism: The temporal dimension of Facebook privacy. In *12th ACM Workshop on Privacy in the Electronic Society*, pages 1–12, 2013.

[3] Paul Bischoff. Two-thirds of secondhand USB drives still contain previous owners' data: study. https://www.comparitech.com/blog/information-security/secondhand-usb-drive-memory-stick-study/, March 2019. Last accessed: 02/2021.

[4] Pew Research Center. Demographics of mobile device ownership and adoption in the United States. https://www.pewresearch.org/internet/fact-sheet/mobile/, Jun 2020. Last accessed 02, 2021.

[5] John Ferguson. How Best Buy's computer-wiping error turned me into an amateur blackhat. https://arstechnica.com/information-technology/2015/06/how-best-buys-computer-wiping-error-turned-me-into-an-amateur-blackhat/, Jun 2015. Last accessed 02, 2021.

[6] Vanessa Forti, Cornelis Peter Balde, Ruediger Kuehr, and Garam Bel. The global e-waste monitor 2020: Quantities, flows and the circular economy potential. http://ewastemonitor.info/wp-content/uploads/2020/12/GEM_2020_def_dec_2020-1.pdf, 2020. Last accessed 02, 2021.

[7] Josh Frantz. Exfiltrating remaining private information from donated devices. https://blog.rapid7.com/2019/03/19/buy-one-device-get-data-free-private-information-remains-on-donated-devices/, Aug 2019. Last accessed 02, 2021.

[8] Simson L. Garfinkel. Carving contiguous and fragmented files with fast object validation. *Digital Investigation*, 4:2–12, 2007.

[9] Simson L. Garfinkel and Abhi Shelat. Remembrance of data passed: a study of disk sanitization practices. *IEEE Security & Privacy*, 1(1):17–27, 2003.

[10] Johannes Götzfried and Tilo Müller. Mutual authentication and trust bootstrapping towards secure disk encryption. *ACM Transactions on Information and System Security (TISSEC)*, 17(2):1–23, 2014.

[11] Grand View Research. Consumer Electronics Market Report for Personal Electronics Industry. https://www.grandviewresearch.com/industry-analysis/personal-consumer-electronics-market, 2020. Last accessed June, 2020.

[12] Blancco Technology Group. Dban help center, Feb 2018. Last accessed 02, 2021.

[13] Hana Habib, Yixin Zou, Aditi Jannu, Neha Sridhar, Chelse Swoopes, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. An empirical analysis of data deletion and opt-out choices on 150 websites. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, 2019.

[14] IDC. Worldwide market for used smartphones forecast to grow to 332.9 million units. https://www.idc.com/getdoc.jsp?containerId=prUS45865720, Jan 2020. Last accessed 02, 2021.

[15] Andrew Jones, Olga Angelopoulou, and L. Noriega. Survey of data remaining on second hand memory cards in the UK. *Computers & Security*, 84:239–243, 2019.

[16] Richard Kissel, Matthew A Scholl, Steven Skolochenko, and Xing Li. Sp 800-88 rev. 1. guidelines for media sanitization, 2006.

[17] Ambar Murillo, Andreas Kramm, Sebastian Schnorf, and Alexander De Luca. "If I press delete, it's gone" — user understanding of online data deletion and expiration. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 329–339, 2018.

[18] Kopo Marvin Ramokapane, Awais Rashid, and Jose Miguel Such. "I feel stupid I can't delete...": A study of users' cloud deletion practices and coping strategies. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, pages 241–256, 2017.

[19] Joel Reardon, David Basin, and Srdjan Capkun. SoK: Secure data deletion. In *2013 IEEE Symposium on Security and Privacy*, pages 301–315. IEEE, 2013.

[20] Joel Reardon, David Basin, and Srdjan Capkun. On secure data deletion. *IEEE Security & Privacy*, 12(3):37–44, 2014.

[21] Linda Serges and Hyla Mobile Inc. Q3 2020 mobile trade-in data. https://blog.hylamobile.com/q3-2020-mobile-trade-in-data, Nov 2020. Last accessed 02, 2021.

[22] Junliang Shu, Yuanyuan Zhang, Juanru Li, Bodong Li, and Dawu Gu. Why data deletion fails? A study on deletion flaws and data remanence in Android systems. *ACM Transactions on Embedded Computing Systems (TECS)*, 16(2):1–22, 2017.

[23] Avast Software. Selling your smartphone could mean selling your identity. Avast finds used smartphones still contain personal information and data. https://press.avast.com/selling-your-smartphone-could-mean-selling-your-identity-avast-finds-used-smartphones-still-contain-personal-information-and-data, Feb 2016. Last accessed 02, 2021.

[24] Android Support. Reset your Android device to factory settings. https://support.google.com/android/answer/6088915. Last accessed: 02/2021.

[25] Apple Support. Erase iPhone. https://support.apple.com/en-ca/guide/iphone/iph7a2a9399b/ios. Last accessed: 02/2021.

[26] Samsung Galaxy Support. Perform a factory reset on your Galaxy phone. https://www.samsung.com/ca/support/mobile-devices/galaxy-phone-perform-a-factory-reset/. Last accessed: 02/2021.

[27] Windows Support. Recovery options in Windows 10. https://support.microsoft.com/en-us/windows/recovery-options-in-windows-10-31ce2444-7de3-818c-d626-e3b5a3024da5. Last accessed: 02/2021.

[28] Michael Yung Chung Wei, Laura M. Grupp, Frederick E. Spada, and Steven Swanson. Reliably erasing data from flash-based solid state drives. In *USENIX Conference on File and Storage Technologies*, 2011.

[29] Davey Winder. Researchers find 'dangerous levels' of sensitive data for sale on eBay. https://www.forbes.com/sites/daveywinder/2019/04/25/researchers-find-dangerous-levels-of-sensitive-data-for-sale-on-ebay/, April 2019. Last accessed: 02/2021.

A Appendices

A.1 Recruitment Advertisement

The following text was used to recruit participants from Facebook Marketplace, Kijiji, and local sub-reddits.
Title: Help with a research study and earn $2 for your participation
Body: Researchers from the University of Guelph are looking for participants for a study on understanding threats to personal data in the second-hand economy. You are eligible to participate in this study if you:

- are 18 years or older
- have currently listed an item for sale on online buying and selling market place like Facebook Marketplace or Kijiji

The study will be performed as an online survey that will take approximately 10 minutes and you will receive $2 for your participation. This research has received ethics approval (Research Ethics Approval Number 19-06-009). If you are interested in participating, please send an email to jceci@uoguelph.ca

A.2 Online Survey

The following multiple choice questions were asked during the online survey.
Demographic Information.

1 How old are you?
(a) 18-25 years old; (b) 26-30 years old; (c) 31-35 years old; (d) 36-40 years old; (e) 41-45 years old; (f) 46-50 years old; (g) Over 50 years old; (h) Choose not to respond

2 What is your gender?
(a) Woman; (b) Man; (c) My gender identity is not listed above; (d) Choose not to respond

3 Which of the following best describes your level of proficiency with technology like smartphones or laptops?
(a) Basic (I can perform basic tasks such as sending emails or browsing the internet);
(b) Intermediate (I can perform intermediate tasks such as changing the settings or installing new applications);
(c) Advanced (I am capable of writing source code)

Electronic Device Lifecycle. For this study, electronic devices refers specifically to electronic devices that can hold personal data such as cell phones, smartphones, laptops, desktop computers, tablets, portable hard drives/flash drives, memory cards and cameras.

4 When you no longer use an old device for any reason, what actions have you taken with your old device? (choose all that apply) (*Each of the following action is asked for the following device types: Smartphones, Laptops, Tablets, Cameras and memory cards, and Hard drives/flash drives*.)
(a) Sell; (b) Give to friend/family; (c) Return/exchange to provider/IT department; (d) Recycle; (e) Throw in garbage; (f) Donate; (g) Keep; (h) Other; (i) I have never stopped using a device of this type.

5 You have answered other to one or more of the above. Please list the other actions you have done with old devices below. (*A box for free form text input is provided*)

6 On a scale from 1 to 5, how concerned would you be if someone (who was not a trusted friend or family member) was able to retrieve the data off of your old devices? (*5-point Likert scale "Least Concerned" - "Most Concerned"*)

Personal Data and Old Devices.

7 **[IF sold, donated, recycled or returned an old device]** You previously answered that you have sold, donated, recycled or returned an old device. Did you remove or attempt to remove your personal data before selling it? (If you have sold, donated, recycled or returned multiple devices, answer for the most recent.)
(a) No, I trust the recipient;(b) No, I am not concerned about the personal data on the device; (c) No, the device is encrypted so my personal data is safe; (d) Yes, I did a "factory reset" or "erase all content"; (e) Yes, I did a zero-fill or secure erase; (f) Yes, I deleted all data manually (i.e., deleting all files); (g) Yes, I deleted some sensitive data.

8 **[IF disposed a device]** How likely do you believe it would be for a person with AVERAGE computer/IT skills to be able to recover any personal data from the device you sold, donated, recycled or returned? (If you have sold, donated, recycled or returned multiple devices, answer for the most recent.) (*7-point Likert scale "Extremely Likely" - "Extremely Unlikely"*)

9 **[IF disposed a device]** How likely do you believe it would be for a person with EXPERT computer/IT skills to be able to recover any personal data from the device you sold? (If you have sold, donated, recycled or returned multiple devices in the same category, answer for the most recent.) (*7-point Likert scale "Extremely Likely" - "Extremely Unlikely"*)

10 **[IF kept an old device]** You previously answered that you kept an old device, what was the main reason you kept the device? (if you have kept multiple devices, answer for the most recent.)

(a) Not worth the hassle of selling or donating; (b) Privacy concerns; (c) Kept it as a backup; (d) Other (*A box for free form text input is provided to specify other*)

A.3 Semi-Structured Interview

The semi-structured interview contained both multiple choice questions (an extension of the online survey) and free form responses to questions that further explored the responses of the participants.

Additional Demographic Information. First, I am going to ask you some additional demographic information about yourself.

1 Which of the following best describes your HIGHEST level of education?
(a) Some high school; (b) Completed high school; (c) Some college/university; (e) Apprenticeship training and trades; (f) Completed college/university; (g) Some graduate education; ; (h) Completed graduate education; (i) Professional degrees; (j) Choose not to answer

2 Which of the following best represents your annual household income?
(a) Less than $30,000 ; (b) Between $30,000 and $74,999; (c) Between $75,000 and $99,999; (d) Over $100,000; (e) Choose not to answer

3 Which of the following best describes your level of proficiency with technology like smartphones or laptops?
(a) Basic (I can perform basic tasks such as sending emails or browsing the internet);
(b) Intermediate (I can perform intermediate tasks such as changing the settings or installing new applications);
(c) Advanced (I am capable of writing source code)

Electronic Device Lifecycle. Next, I am going to ask you about what you do with old devices you no longer use. For this study, electronic devices refers specifically to electronic devices that can hold personal data such as cell phones, smartphones, laptops, desktop computers, tablets, portable hard drives / flash drives, memory card and cameras.

4 What type of data do your old devices contain or previously contained? (choose all that apply) (*Each of the following action is asked for the following device types: Smartphones, Laptops, Tablets, Cameras and memory cards, and Hard drives/flash drives*.)
(a) N/A; (b) Contacts; (c) Emails; (d) Photos; (e) Adult Content; (f) Personal adult content (myself or partner); (g) banking info / credit card; (h) Text and instant messages; (i) Passwords; (j) Personal videos; (k) Browser history.

5 On a scale from 1 to 5, how concerned would you be if someone was able to retrieve the data off of your old devices? (choose all that apply; choose N/A for devices you have not owned) (*5-point Likert scale "Least Concerned" - "Most Concerned"; User provides answer for each of the following categories: smartphones, laptops, tablets, cameras and memory cards, hard drives/flash drives*)

[IF gave a device away] Giving Devices Away.

6 You previously answered that you have given an old device to a friend or family member. Did you attempt to remove your personal data before giving it to them? (If you have given away multiple devices in the same category, answer for the most recent.) (*Each of the following action is asked for the following device types: Smartphones, Laptops, Tablets, Cameras and memory cards, and Hard drives/flash drives*.)
(a) No, I trust the recipient;(b) No, I am not concerned about the personal data on the device; (c) No, the device is encrypted so my personal data is safe; (d) Yes, I did a "factory reset" or "erase all content"; (e) Yes, I did a zero-fill or secure erase; (f) Yes, I deleted all data manually (i.e., deleting all files); (g) Yes, I deleted some sensitive data; (h) N/A.

7 If response was yes to previous question, researcher asked how was the data erased or how was the device wiped or reset. If applicable, researcher asked participants to tell the steps taken for each device type and ask why participants used a certain method.

[IF sold a device] Selling Old Devices.

8 You previously answered that you have sold an old device. Did you remove or attempt to remove your personal data before selling it? (If you have sold multiple devices in the same category, answer for the most recent.) (*Each of the following action is asked for the following device types: Smartphones, Laptops, Tablets, Cameras and memory cards, and Hard drives/flash drives*.)
(a) No, I trust the recipient;(b) No, I am not concerned about the personal data on the device; (c) No, the device is encrypted so my personal data is safe; (d) Yes, I did a "factory reset" or "erase all content"; (e) Yes, I did a zero-fill or secure erase; (f) Yes, I deleted all data manually (i.e., deleting all files); (g) Yes, I deleted some sensitive data; (h) N/A.

9 If response was yes to previous question, researcher asked how was the data erased or how was the device wiped or reset. If applicable, researcher asked participants to tell the steps taken for each device type and ask why participants used a certain method.

10 How likely do you believe it would be for a person with AVERAGE computer/IT skills could recover any personal data from the device you sold? (If you have sold

multiple devices in the same category, answer for the most recent.) (*7-point Likert scale "Extremely Likely" - "Extremely Unlikely"*)

11 How likely do you believe it would be for a person with EXPERT computer/IT skills could recover any personal data from the device you sold? (If you have sold multiple devices in the same category, answer for the most recent.) (*7-point Likert scale "Extremely Likely" - "Extremely Unlikely"*)

12 You previously answered that you have sold an old device in the past, which personal data category would you be most worried about a buyer accessing? (If you have sold multiple devices in the same category, answer for the most recent.)
(a) N/A; (b) Contacts; (c) Emails; (d) Photos; (e) Adult Content; (f) Personal adult content (myself or partner); (g) banking info / credit card; (h) Text and instant messages; (i) Passwords; (j) Personal videos; (k) Browser history.

[IF donated or recycled a device] Donating or Recycling Devices.

13 You previously answered that you have donated or recycled at least one old device. Did you remove or attempt to remove your personal data before donating or recycling it? (If you have sold multiple devices in the same category, answer for the most recent. Answer N/A if you have not donated or recycled a device of that type.) (*Each of the following action is asked for the following device types: Smartphones, Laptops, Tablets, Cameras and memory cards, and Hard drives/flash drives*.)
(a) No, I trust the recipient;(b) No, I am not concerned about the personal data on the device; (c) No, the device is encrypted so my personal data is safe; (d) Yes, I did a "factory reset" or "erase all content"; (e) Yes, I did a zero-fill or secure erase; (f) Yes, I deleted all data manually (i.e., deleting all files); (g) Yes, I deleted some sensitive data; (h) N/A.

14 If response was yes to previous question, researcher asked how was the data erased or how was the device wiped or reset. If applicable, researcher asked participants to tell the steps taken for each device type and ask why participants used a certain method.

[IF returned a device] Returned Devices.

15 You previously answered that you have return at least one old device to the provides, IT department or manufacturer. Did you remove or attempt to remove your personal data before returning it? (If you have returned multiple devices in the same category, answer for the most recent. Answer N/A if you have not donated or recycled a device of that type.) (*Each of the following action is asked for the following device types: Smartphones, Laptops, Tablets, Cameras and memory cards, and Hard drives/flash drives*.)
(a) No, I trust the recipient;(b) No, I am not concerned about the personal data on the device; (c) No, the device is encrypted so my personal data is safe; (d) Yes, I did a "factory reset" or "erase all content"; (e) Yes, I did a zero-fill or secure erase; (f) Yes, I deleted all data manually (i.e., deleting all files); (g) Yes, I deleted some sensitive data; (h) N/A.

16 If response was yes to previous question, researcher asked how was the data erased or how was the device wiped or reset. If applicable, researcher asked participants to tell the steps taken for each device type and ask why participants used a certain method.

[IF disposed a non-functioning device] Non-Functioning Devices.

17 For each device type, which of the following best describes what you have done with a **non-functioning** (broken, damaged) device you no longer use? (If you have had multiple non-functioning devices in the same category, answer for the most recent. Answer N/A if you do not have not had a non-functioning device of that category.) (*Each of the following action is asked for the following device types: Smartphones, Laptops, Tablets, Cameras and memory cards, and Hard drives/flash drives*.)
(a) Donate or recycle it; (b) Destroy it; (c) Throw it in the trash; (d) keep it; (e) Sell or give it away; (f) N/A.

18 You have previously answered that you have donated, recycled, sold, or given away a non-functioning device before, did you attempt to remove your personal data first? (If you have had multiple non-functioning devices in the same category, answer for the most recent. Answer N/A if you do not have not had a non-functioning device of that category.)
(a) No, I believe it would require too much effort for someone to retrieve my personal data; (b) No, my device was encrypted; (c) No, I haven't considered my personal data privacy in this case; (d) No, I don't know how to or it was too difficult to remove my personal data; (e) Yes, I removed or attempted to remove my personal data.

19 You previously answered that you have thrown away a non-functioning device before, did you attempt to remove your personal data first? (If you have had multiple non-functioning devices in the same category, answer for the most recent. Answer N/A if you do not have not had a non-functioning device of that category.)
(a) No, I believe it would require too much effort for someone to retrieve my personal data; (b) No, my device

was encrypted; (c) No, I haven't considered my personal data privacy in this case; (d) No, I don't know how to or it was too difficult to remove my personal data; (e) Yes, I removed or attempted to remove my personal data.

20 If response was yes to previous question, researcher asked how was the data erased.

Personal Data on Purchased Devices.

21 Have you ever purchased an electronic device that had personal data from the previous owner/user?
(a) Yes; (b) No; (c) Maybe

22 If response to the last question is yes, the researcher asked what type of data did they find from the previous owner? What device or device type was it? (iPhone, laptop, etc.)

Protecting Personal Data.

23 How difficult do you currently believe it is to fully remove all personal data from the following devices? (*Each of the following action is asked for the following device types: Smartphones, Laptops, Tablets, Cameras and memory cards, and Hard drives/flash drives*.)
(a) Easy; (b) Intermediate; (c) Difficult or Impossible; (d) N/A

24 I think the following is true about the "Reset Device" feature on my laptop or personal computer:
(a) It does not provide information that assists me in selling the device to a stranger; (b) It provides information that assists me in selling the device to a stranger; (c) N/A or Unknown

25 I think the following is true about the "Reset Device" feature on my smartphone:
(a) It does not provide information that assists me in selling the device to a stranger; (b) It provides information that assists me in selling the device to a stranger; (c) N/A or Unknown

26 When users are selling a used electronic device, I feel it is the responsibility of the following entities to inform users about the threats to personal data on the device: (*4-point Likert scale responses"Strong responsibility", "Some responsibility", "No responsibility but should assist", "No responsibility"*)
(a) Online platform/store (e.g., Kijiji or eBay); (b) Sellers themselves; (c) Device manufacturers (e.g., Apple or Samsung)

27 Has the difficulty to remove data from a device ever made you reluctant to sell or donate that device?
(a) Yes; (b) No

28 I believe that classifieds/online marketplaces should explain the risks associated with selling used devices and how to properly wipe used devices. (*5-point Likert scale "Strong agree" - "Strongly disagree"*)

29 I believe that device manufacturers should make it easier to securely remove all personal data from electronic devices. (*5-point Likert scale "Strong agree" - "Strongly disagree"*)

30 If an electronic device had a feature to securely remove all personal data, I would be more likely to purchase that device compared to a device that did not. (*5-point Likert scale "Strong agree" - "Strongly disagree"*)

31 If an electronic device had a feature to securely remove all personal data, I would be more likely to sell the device when I no longer required it. (*5-point Likert scale "Strong agree" - "Strongly disagree"*)

32 The researcher asked participants to rank the following terms in regards to how effective they are at removing and preventing recovery of personal data from a computer or electronic device before selling or recycling it: "Clean the drive", "Delete all files", "Erase the hard drive", "Secure erase the hard drive". (If two choices are equally effective, they were asked to assign them the same rank. The researchers noted the confusions that participants had regarding the use of these terms.)

A.4 Detailed Interview Findings

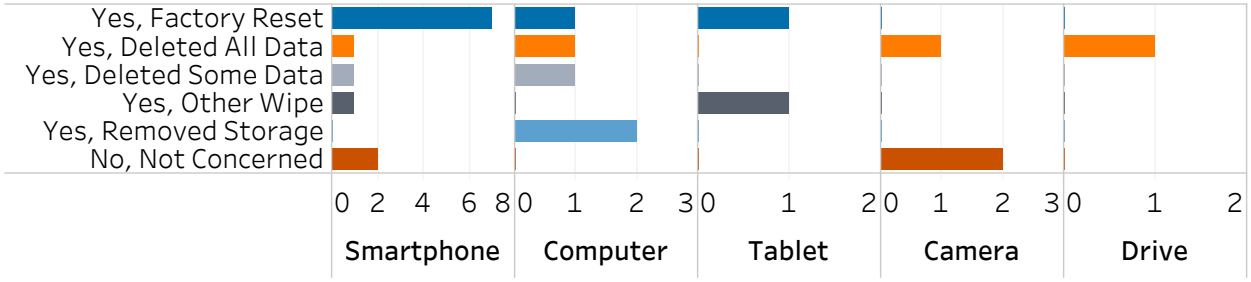


Figure 9: Participants' responses to "Did you remove any data from the device before donating or recycling it?"

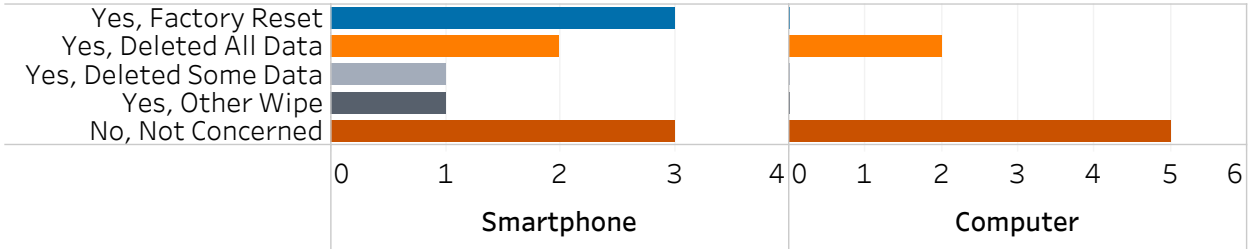


Figure 10: Participants' responses to "Did you remove any data from the device before returning it to a provider, IT department, manufacturer or retailer?"

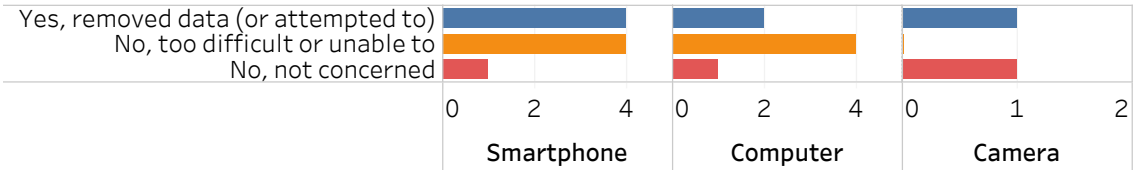


Figure 11: Participants' responses to "Before you sold, gave away, returned, threw away, recycled or donated the broken device, did you remove any data from it?"

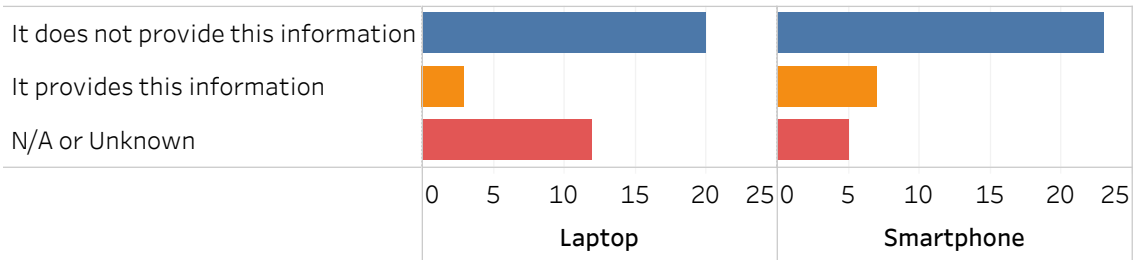


Figure 12: Participants' responses to "If you have used the reset feature in the past, did it provide you with any information that assists you when selling the device to a stranger? For example, if information is recoverable."

Exploring Authentication for Security-Sensitive Tasks on Smart Home Voice Assistants

Alexander Ponticello
CISPA Helmholtz Center for Information Security and Saarland University

Matthias Fassl
CISPA Helmholtz Center for Information Security and Saarland University

Katharina Krombholz
CISPA Helmholtz Center for Information Security

Abstract

Smart home assistants such as Amazon Alexa and Google Home are primarily used for day-to-day tasks like checking the weather or controlling other IoT devices. Security-sensitive use cases such as online banking and voice-controlled door locks are already available and are expected to become more popular in the future.

However, the current state-of-the-art authentication for smart home assistants consists of users saying low-security PINs aloud, which does not meet the security requirements of security-sensitive tasks. Therefore, we explore the design space for future authentication mechanisms.

We conducted semi-structured interviews with $N = 16$ Alexa-users incorporating four high-risk scenarios. Using these scenarios, we explored perceived risks, mitigation strategies, and design-aspects to create secure experiences. Among other things, we found that participants are primarily concerned about eavesdropping bystanders, do not trust voice-based PINs, and would prefer trustworthy voice recognition. Our results also suggest that they have context-dependent (location and bystanders) requirements for smart home assistant authentication. Based on our findings, we construct design recommendations to inform the design of future authentication mechanisms.

1 Introduction

Voice-controlled smart home assistants find their way into more households every year. Gartner estimates that, by the year 2025, half of the knowledge workers will use voice assistants every day [10]. Currently, voice assistants offer entertainment (e.g., playing music, games), information gathering (e.g., weather, cooking recipes), and personal planning (e.g., calendar, task list). They are also a control hub for smart home IoT devices, such as smart light bulbs or heating. However, vendors already work towards new and more security-sensitive use cases for these assistants. Voice-based online shopping allows users to order goods without interrupting their current activity. Compatible locking systems permit users to open doors via voice commands [7]. Capital One, a technology-focused bank in the U.S., uses the Amazon Alexa platform to offer bank services such as retrieving account information, including their current balance, or paying credit card bills [12].

However, as Abdi et al. [2] found, security and privacy concerns hinder user adoption of these new use cases for voice assistants. Amazon Alexa, a widespread voice assistant that supports online shopping, currently only offers an optional voice code to authenticate users before their purchase. This simplistic authentication method is insufficient for more security- and privacy-critical tasks. Hence, voice assistants need more robust protection mechanisms. Our community already invested a significant effort in developing and improving authentication mechanisms for various tools and use cases [9, 14, 20]. However, designing authentication for voice assistants comes with unique challenges since they usually do not offer I/O methods beyond the voice channel. This limitation makes transferring existing authentication mechanisms to voice assistants difficult. Hence, we need device-appropriate authentication mechanisms for voice assistants. Developing these starts with finding all viable forms of authentication that users trust.

In this work, we explore the design space of authentication with voice assistants in a user-centered way. We conducted semi-structured interviews with $N = 16$ participants that included four scenarios. These scenarios depicted different situations in which the protagonists perform security-sensitive tasks with a voice assistant. We evaluated the transcribed in-

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

terviews using Thematic Analysis [11] to explore the design space. Our contribution includes findings on: (1) users' perception of threats, (2) users' mitigation strategies in security-sensitive circumstances, (3) users' expectations for authenticating with voice assistants, and (4) implications for the design of future authentication mechanisms. Our results show that users see bystanders in hearing range as a potential threat to their security and privacy. Their main mitigation responses focus on limiting their use of security-sensitive features. Hence, developing alternative user-trusted authentication mechanisms is crucial to facilitate adoption of security-sensitive use cases. The participants appreciated the low-effort interaction with voice assistants and expected similar from authentication. Voice-based biometric authentication fulfills that criterion and was frequently suggested for authentication. In social situations, participants reported discomfort with voice code authentication and privacy-sensitive tasks. Hence, an additional *discreet mode* for voice assistants potentially improves adoption rates. Participants described that their trust in security mechanisms builds with experience. Therefore, we suggest that voice assistants provide a *demonstration mode* for security- and privacy-related features.

2 Related Work

Our work builds on different areas of prior work: security and privacy of smart home environments as well as voice assistants, alternative authentication schemes for voice user interfaces, and users' risk perceptions and mitigation strategies.

2.1 Security and Privacy of Smart Homes

Zeng et al. [39] studied users' mental models of smart home systems and threats. They found incomplete mental models of how IoT devices, including voice assistants, interact with each other and with back-end cloud services. Building upon this work, Zeng and Roesner [40] explored users' security and privacy issues in a month-long in-home study. They identified several open challenges, most importantly incorporating voice assistants into access control systems so that they can become effective control hubs for smart homes. In this context, they highlight the importance of sophisticated voice-based authentication, which motivates our study.

Yao et al. [37] conducted a co-design study with users and non-users of smart home technologies to investigate their privacy concerns and needs. They identified key design factors for smart home privacy controls, including authentication for multiple users and access control.

Zimmermann et al. [43] studied potential users' mental models of smart homes. Their participants had sparse mental models of smart home systems, and almost all of them were concerned about their personal data's security.

Yao et al. [38] used three scenarios to study bystanders' privacy perceptions in smart homes, i.e., people living in or visiting smart homes where they are not primary users. Bystanders were concerned about the video and audio data collection and demanded privacy controls tailored to them specifically. Furthermore, the authors highlight how the users' role in smart homes, e.g., system owners and bystanders, can strain their relationship.

2.2 Security and Privacy of Voice Assistants

Huang et al. [18] examined privacy perceptions and coping strategies of users sharing voice assistants. They found that users with limited mental models did not understand how the system shares their data with other users. In contrast, participants with more advanced mental models were concerned about the immature technology, e.g., voice recognition to distinguish users. The authors highlight the need for more sophisticated authentication mechanisms to tackle these issues.

Lau et al. [21] conducted a diary study and interviews to shed light on privacy perceptions and privacy-seeking behaviors around voice assistants. They found that voice assistant users did not entirely understand privacy risks and frequently traded privacy for increased convenience. Non-users were concerned about privacy and security, partially because of their limited trust in voice assistants' manufacturers.

Chalhoub and Flechais [13] report similar findings from their qualitative study exploring the effect of user experience (UX) factors on voice assistant users' security and privacy. They found that common security and privacy features, such as muting, were not user-friendly. As a response, users disabled features or disconnected their devices.

Zhang et al. [41] describe the *Dolphin Attack*, a novel technique that utilizes ultrasonic audio signals to inject commands into smart home voice assistants. These commands are inaudible to humans and exploit the non-linearity property of current microphones, which is why they treat high-frequency sounds similar to genuine human speech. The authors bypassed the biometric voice authentication of up-to-date smart home voice assistants by combining this attack with users' resampled legitimate audio snippets. Roy et al. [29] build upon this work, extending the attack range from 1.5m to 7.6m using an array of speakers. Sugawara et al. [34] developed an attack called *LightCommands*. By exploiting a vulnerability in micro-electro-mechanical systems (MEMS) microphones, this technique allows an attacker to inject commands via a potent light source. Since these commands are transmitted by light, attackers can inject them from afar while victims cannot hear them. The authors report a successful command injection over a distance of 75m with a laser beam aimed at a Google Home device behind a glass window. Lei et al. [22] make use of channel state information in Wi-Fi networks to detect human presence in a room. Assuming that most attacks happen during users' absence, VUI systems only ac-

cept commands if someone is present at that time. Related work identified a gap between users' expectations of potential threats and technically feasible attacks. Using our study, we also want to increase our knowledge about this gap and laying the groundwork to reduce it in the future.

2.3 Authentication for Voice User Interfaces (VUIs)

Feng et al. [14] designed a wearable-based authentication scheme for VUI. Their system verifies VUI commands by independently recording voice commands from skin vibrations. Hence, this method provides continuous authentication. They tested different designs for the wearable, such as earbuds, necklaces, and glasses.

Blue et al. [9] proposed a similar scheme using a second microphone-equipped device, e.g., a smartphone. By measuring the direction of arrival of each voice command, their system can detect whether the speaker is closer to the VUI or the second microphone. Assuming that users carry their smartphone on them during VUI interaction, the system only deems nearby commands authentic.

Kwak et al. [20] employed machine learning to differentiate genuine user commands from malicious input.

Zhang et al. [42] developed a system for speaker liveness detection. By extracting features in the Doppler shifts, they can distinguish audio generated by an artificial speaker from a human voice. Their system enhances voice authentication by protecting from common threats, e.g., *Replay Attacks*.

The presented authentication systems build upon assumptions about how users interact with VUIs and how they perceive the system. In this work, we use qualitative methods to explore the underlying design space, thereby laying the groundwork for future authentication systems taking users' security and privacy needs into account.

2.4 Users' Risk Perceptions and Mitigation Strategies

Several other works used methodological approaches similar to this paper's to investigate users' risk perceptions and mitigation strategies outside of the smart home context. Many of their findings are observable across various systems and technologies. Hence, they potentially apply to voice assistants as well.

Harbach et al. [16] studied Internet users' risk awareness. The results indicate that most of the 210 participants were aware of general risks. The authors state that users are aware of seven risks on average, which significantly vary across persons, populations, and the interaction's context. They highlight that existing security measures often focus on technical risks of which users are less aware. Since users have a limited compliance budget, the authors argue that they might not adopt measures that do not directly address their perceived relevant risks. Furthermore, the authors propose improving risk communication and education to support the users' risk perception.

Ruoti et al. [30] conducted interviews with middle-aged suburban parents about their online security posture. They found that users weigh the trade-offs between gained security and necessary effort when choosing security mechanisms. Due to participants' perception that complete security is unobtainable, they less frequently adopt cost-intensive coping strategies. They identified a four-step process users pass through where they first learn about a new security threat (e.g., by news reports), evaluate their personal risk (ignoring threats perceived as unlikely), estimate the damage in terms of the effort they have to invest after a breach, and selecting an appropriate coping strategy after weighing the costs and benefits.

Stobert and Biddle [33] studied coping strategies that users apply when managing passwords. They found that some of the most prominent mitigation strategies, e.g., writing down passwords or reusing them, seem to disregard popular security advice. The authors argue that this behavior is not caused by users' insufficient risk perceptions but rather that users make rational choices based on their personal resources.

3 Methodology

We chose semi-structured interviews building on previous works [8, 39] to answer the following research questions:

RQ1 Which attackers and threats are users concerned about when performing high-risk tasks via voice-controlled assistants in a smart home environment?

RQ2 Which potential mitigation strategies do users apply to protect themselves?

RQ3 Which properties does an authentication system for voice assistants need such that users perceive it as secure?

3.1 Procedure

We briefed participants on the topic and purpose of the study and how data is processed and handled. All participants signed consent forms that permitted audio recordings. Our interview guideline (presented in Section A) consists of three parts. First, we asked a series of warm-up questions regarding our participants' general Alexa usage and experiences with the online shopping feature.

In the second part of our interview guideline, we presented participants with four scenarios on *vignettes*. These included pictures and short textual descriptions of an interaction between a user and their smart home assistant. Prior work showed that scenarios are a useful tool for examining users'

perceptions and mental models of a system [2, 4, 19, 36]. Vignettes allowed participants to immerse themselves into situations, which would have been more difficult using interview questions alone. Vignettes are closer to reality than abstract questions and might reduce social desirability bias by allowing interviewers to ask questions less directly [26].

Scenarios The scenarios combined four security-sensitive tasks with different situations. These situations vary in two aspects: the number of bystanders and the location, namely inside or outside the house. Most of the presented functions are currently not available in central Europe. The scenarios are as follows:

- **Dinner.** This scenario combined the task of transferring a small amount of money during a dinner party with friends. The use of Alexa can be convenient since the user is sitting at a table. Several bystanders might eavesdrop on the interaction. However, these people are, to a certain degree, trustworthy as they are close acquaintances. The corresponding image shows a laid table with several people around, chatting in a light atmosphere.
- **TV.** This scenario involves users and their partners. We selected the activity of paying a reoccurring bill since this is a typical task concerning both partners while also being less casual and less frequent. The picture associated with this scenario depicts two people sitting in a living room on a couch in front of a running TV.
- **Door.** We combined the task of unlocking the front door with the scenario of coming back from grocery shopping. A typical task that users perform while outside the house. The situation includes the user carrying several bags, making unlocking doors more difficult. This setting justifies the use of a voice-controlled smart home assistant. No other people are immediately present in the scene. The picture shows a person carrying bags of groceries next to a car, a blue sky in the background indicates that the scene takes place outside.
- **Hands.** We coupled the task of checking a transaction history with gardening work, making the protagonist's hands dirty. We included children as potential bystanders. We described them as running around and screaming, meaning they do not pay immediate attention to the user while still being present. Also, this scenario does not feature a dialog with Amazon Alexa. In this description, we do not refer to Alexa nor include a device in the image to leave room for the interviewee to imagine how an interaction could play out. At the same time, this allows us to explore alternative interaction mechanisms, potentially not involving Alexa.

We presented the scenarios in random order. During on-site interviews, we presented the vignettes on printed and laminated cards face-down to participants. For remote interviews, we showed participants a website that displayed four face-down cards. In both cases, we flipped and discussed the cards in the participants' chosen order. Section C presents the full vignettes that we used for the interviews. We provided pen and paper for note-taking to participants or asked them to send us their drawings by email during remote interviews respectively.

After letting them read the description text and look at the image, we asked participants which problems they think could arise in such a situation. We did not ask about security-related problems to avoid priming participants in a specific direction. If participants mentioned no security-related problems, we followed up with respective questions, e.g., whether they thought the voice code included in the scenario was useful or not. Then, we explored threats that participants identified and asked them to think of any other actors posing threats and their potential mitigation strategies. We investigated what interviewees thought might be useful to them and which mitigation strategies they would apply in the given scenario, with the threats described above in mind. We repeated this process for all four scenarios. Afterward, we asked participants to summarize all four situations and to think about possible similarities and differences between the scenarios, possibly applying the insights they gained in a later scenario to an earlier one. This recapitulation also helps to focus participants on details they might not have noticed before (e.g., the number of people present in the scenario) and think about the consequences introduced by said factors.

In the third and final part of our interview guideline, we included some demographic questions, mostly used to describe our sample. We included two standardized scales in this section, namely the ATI scale [15] and the CFIP scale [32].

After each interview, we asked the participant about any remaining questions, reiterated the study's purpose, and explained why we designed the interview guideline and the vignettes as presented. We also explained the current situation regarding security, and most of all, authentication, on Amazon Alexa and comparable VUIs.

Accessibility One blind Alexa user participated in our study. We adapted our study material to ensure accessibility and exchanged the printed vignette cards with two separate audio recordings: To have a clear distinction between vignette descriptions and interview questions, one author, who was not the interviewer, narrated the picture displayed on top of the card. In a separate audio file, the narrator read the corresponding text aloud. We used a computer-generated voice similar to the one from Alexa to illustrate interactions with the voice assistant. Providing two recordings allowed the participant to replay each part separately.

Pilot Interviews We pilot-tested our interview guideline with two on-site and two remote interviews. Based on the

findings, we decided how much time to allocate as well as financial compensation. We dropped one interview question as it was too ambiguous; we also modified the presentation of the *door* vignette to clarify the scenario. We excluded the pilot interviews from the final dataset.

3.2 Thematic Analysis

We transcribed the data at an orthographic level, including non-verbal utterances only when we deemed them essential for the semantic of a phrase (e.g., a participant laughing while saying something, indicating it was a joke). Afterward, we read and re-read the data to get an even better understanding, taking notes of interesting details and basic patterns.

We chose to analyze our data using a thematic analysis approach, as described by Braun and Clarke [11]. We conducted the interviews in English and German, coded the resulting data in German, and later translated the codebook to English.

To construct a codebook, two researchers performed open and axial coding on a subset of four interviews. First, we performed open coding, then met to resolve disagreements and re-coded the data. Krippendorff's alpha was 0.50 before and 0.94 after the discussion and re-coding step, indicating a high agreement. Then, we performed axial coding (on the same subset of interviews) to identify higher-level themes. Then, another subset of four interviews was coded with a Krippendorff's alpha of 0.83, indicating a strong agreement between the two coders. At this stage, the existing codebook covered most of the data's aspects, so we only sparingly introduced new codes. Finally, one researcher coded the remaining interviews using the codebook agreed upon in the previous discussion.

3.3 Recruitment and Participants

We mainly recruited Alexa users because Alexa has the largest share of the smart speaker market, and its (security-sensitive) shopping feature is well-developed and widespread. However, we also welcomed participants who had experience with other types of voice assistants.

In total, we recruited 16 participants in Germany (9), Austria (6), and Italy (1); five of them via flyers around our institution's campus, three participants over mailing lists; six via convenience sampling; and two via snowball sampling. We stopped recruiting new participants after we reached saturation for our target population, i.e., Amazon Alexa users from Central Europe without computer science background. Due to the COVID-19 pandemic, we conducted ten interviews online and six in person at our department. We compensated all participants with a 15 Euro Amazon voucher, which is in line with similar studies [19, 39].

In total, we recruited seven women and nine men. Their average age was 29.31 ($\sigma = 10.69$, median $= 26.5$). Fourteen participants had at least completed high school, with seven holding a bachelor's degree (or equivalent), and two holding a master's degree (or equivalent). We also measured the participants' affinity for technology interaction using the seven-point ATI scale [1]. The average ATI score was 4.1 ($\sigma = 0.76$, median $= 3.83$), which is above the population-wide average of 3.5. To assess people's privacy concerns we used the seven-point CFIP scale [17]. The average CFIP score was 5.76 ($\sigma = 0.71$, median $= 5.93$).

3.4 Ethical Considerations

Our institution's ethical review board (ERB) reviewed and approved our study. We followed our principle of minimizing the collection of personally identifiable information (PII) as far as possible. We stored and processed data in line with the GDPR and our institution's ethical regulations. We collected informed consent from all participants and informed them how we would process their data. If participants had further questions or wished to withdraw their consent afterward, they could use the provided contact information.

3.5 Positionality Statement and Expectations

In the spirit of constructivism, we assume that our personal views as researchers shape every part of a study, from study design to data analysis to reporting. Here, we want to make our a priori expectations (similar to Krombholz et al. [19] and Braun and Clarke [11]) transparent. We focus on the expectations that influenced the design of the four scenarios.

We expect that the presence of bystanders (esp. considering the familiarity between the user and the bystander), the location in which users perform a task, and the task's perceived security-sensitivity (esp. considering financial risks or potential risk of a property's physical security) are most likely to influence the participants' responses.

E.g., we expect users to neglect the threat of other IoT devices listening in on their actions but hypothesize that they perceive bystanders as potential risks. Furthermore, we expect that users have incorrect assumptions about the security of authentication methods and the kind of threats they mitigate. Regarding mitigation strategies that users employ, we expect to find that people refrain from using the system entirely or only use security-critical features when bystanders are not present. Some users might use Alexa's whisper mode to prevent other people from overhearing a sensitive conversation.

4 Results

We now present the exploratory findings of the design space for smart home assistant authentication. When analyzing our data, we focused on answering our research questions stated in Section 3. This chapter is structured according to the categories we developed during the axial coding step. First, we cover the perceptions of threats. Users were concerned about

different attackers that could affect them in the presented scenarios. They also reflected on trust in certain groups of people or entities. Next, we report mitigation strategies that participants considered to protect themselves. These mitigation strategies improve our understanding of how participants use these systems and which practices they adopt to mitigate threats. Finally, we present essential properties for secure and usable smart home assistant authentication we discovered during our data analysis.

For easier readability, we refer to individual participants with labels P1-16 throughout this section.

4.1 Concerns about Attackers and Threats

We answer RQ1 by reporting perceptions of threats and attackers that users were concerned about when performing security-sensitive tasks on a voice assistant. We found that most users perceived bystanders as potential threats. Both familiar (e.g., family, friends) and less familiar (e.g., neighbors, casual visitors) bystanders could be present during an interaction with Alexa, meaning that the voice code used for authentication could be eavesdropped on by an intentional attacker or an accidental listener.

Insiders We discovered several conflicting perceptions about insiders as a threat. Similar to previous work [18, 24], almost all participants agreed that they trust their friends in general, however, we found that this does not always extend to security- and privacy-related affairs. P7 states: *"I trust my friends, but not with my money."* Correspondingly, most interviewees showed a more extensive amount of trust towards a partner, some of them even willingly sharing their authentication code. Others, however, expressed concerns that a partner might become a threat if the relationship were to end on bad terms. Previous work by Levy et al. [23] and Marques et al. [25] suggests widespread adversarial behavior between family members. Lastly, we found the perception that children are a potential threat, depending on their age. P4 explains that: *"Children are usually quite bright and soak everything up like a sponge, and I think they could use that somehow, the voice code, to make transfers or top up their phone."*

When we asked participants about the possible motivation of insider threat actors, they suspected that friends and children would prank them. While such pranks usually do not cause much harm, they present an inconvenience that most participants prefer to avoid.

Criminals Participants also considered more serious attackers such as criminals, both on- and offline, which is in line with results of previous work [2, 16, 18, 30, 43]. In the presented scenarios, criminals could be especially motivated by the potential high financial gains. As also reported by other researchers [39, 43], physical access to their home proved to be a primary and widespread protection measure within our sample. In our specific scenarios, interviewees showed awareness of several attack vectors, namely eavesdropping, *Replay Attacks*, and brute-force attacks on voice codes. Participants expected attackers to employ readily available devices such as microphones to capture a user's interaction with a voice assistant. While most thought such an attack would need to happen in situ, a few interviewees were aware of voice sampling techniques using arbitrary audio of a user to produce adversarial samples.

In the context of personal finance scenarios, participants were concerned about remote attackers interfering with their devices over the Internet. These attackers could exploit Alexa's vulnerabilities to eavesdrop on a user's voice code or inject malicious commands directly, in both cases bypassing authentication. Some interviewees also suspected that attackers use other IoT devices to monitor users and interfere with voice assistants. Similarly, some were aware of malicious skills as potential attack vector.

Untrustworthy or faulty infrastructure Due to past experiences, interviewees expressed concerns about technical issues impeding a secure interaction with Alexa. They highlight that failures of the speech-to-text system might lead to wrong or unauthorized commands getting executed, marking a security breach. These findings add to results by related work, demonstrating how unintended or miss-interpreted voice commands can lead to violations of users' privacy [24] and frustration during authentication [35]. Schönherr et al. [31] found over 1000 triggers for Amazon Alexa, Google Assistant, and other smart home assistants in TV-shows, news, or audiobooks.

Finally, almost all users expressed privacy concerns when it comes to sharing data with Amazon. High-risk tasks such as money transfers can involve sensitive data that participants were uncomfortable sharing with a company they suspected of employing targeted advertisement or selling data to third parties. Storing user data renders data leaks on the backend of the system possible, potentially due to cyberattacks. Finally, some users also explained that Amazon or its employees might eavesdrop on a user's voice code and use it against their will.

4.2 Mitigation Strategies

We address RQ2 by reporting the participants' mitigation strategies. Users largely agreed they would refrain from using an authentication system they perceive as insecure, especially if they consider the use case non-essential. This matches users' coping strategies in other contexts, e.g., online shopping or setting up smart home systems [2, 18, 24, 39]. Our findings suggest that users generally perceive Alexa as a luxury item that facilitates tasks but does not enable previously unavailable features. Hence, using Alexa for security-sensitive tasks is just an additional attack surface for the participants. As P4

phrases it: *"I wouldn't use any of the skills described here because the effort- or the comfort-to-risk ratio is not profitable for me."* We found, similiar to Abdi et al. [2], that users preferred employing personal computers or smartphones as fall-back authentication method. Mainly because users have pre-established trust with these devices.

We found that users employ a trial-and-error strategy to build up trust and improve their understanding of the protection provided by authentication systems. By trying out the system under typical attacking conditions, users could gain trust in a novel mechanism. We found a go-to attack for this technique is mimicking a legitimate user's voice. Users would test voice biometric authentication by *"sit[ting] in front of it quite often and try[ing] it out while disguising my voice, to see whether Alexa still recognizes me or not."* (P10). Most interviewees had not used voice-based authentication before and did not trust a system without hands-on experience. We found that both positive past experiences and a lack of negative ones can give users a sense of security. P1 explains this as follows: *"there may be some [security] issues with the payment method I'm currently using, but I've done it so often and I'm so familiar with it that I feel safer because of that."*

We found that eavesdropping was the users' prime concern. Hence, interviewees presented various mitigation strategies for this threat. The most prominent one was moving to another room if several bystanders were present, e.g., in the scenario *"Dinner"*. Huang et al. [18] reported similar user concerns and coping mechanisms in a less security-sensitive task: making phone calls via voice assistants. We found that there exist specific situations in which users do not desire voice interaction. Some participants stated that this was due to an awkward feeling when talking to a computer, which can be perceived as *"admitting to being lazy"* (P10) because a user does not carry out tasks themselves, delegating them to a computer instead. Furthermore, interaction over voice can draw unwanted attention to the user. Participants expressed a desire for discreet interaction options, especially for money-related tasks; *"money is always a delicate topic and you don't want to address that in front of everyone"* (P4). Using the whisper mode of Alexa can be a less obtrusive operation mode. Participants also stated that this mode potentially mitigates eavesdropping. However, this input feature was perceived as less elegant and, consequently, not fitting into *"the Alexa lifestyle"* (P6).

Another mitigation strategy for eavesdropping was changing the code regularly. By doing so, participants expected that a leaked code would no longer be valid during an attack. Similarly, interviewees described more complex codes as hard to remember and, therefore, also difficult for an eavesdropper to pick-up. We found that users believed they could recognize on-going attacks against their devices while present. P7 notes: *"Inside the house, no real sound can get through. If someone stands in your garden and yells: ALEXA! [. . .] then you probably hear it too."* Therefore, attacks would mainly occur while they were away from home. In this case, participants desired stronger than usual security measures. Our findings suggest that users are not aware of attacks injecting inaudible voice commands, possibly from outside the house [34, 41].

While participants did not perceive voice codes as an adequate authentication mechanism for general use cases, some interviewees talked about its positive effects. A voice code can be an effective mitigation strategy against accidentally executed commands since, unlike regular voice commands, participants did not imagine saying their code in a casual conversation. Some interviewees perceived the code as a minimum security mechanism protecting them from their friends' or children's pranks. They preferred using a code over having no security measures. P9 explains that it *"just gives another layer of security, so my friend couldn't just come into my house and be like: Alexa, pay the utility bill!"* Several participants mentioned remote attackers as a concern, though none could think about mitigation strategies against this threat. Participants did not talk about preventive measures such as keeping systems up to date during the interviews. This observation is in line with Anell et al.'s findings [6].

4.3 Important Properties of Authentication Systems

We present important aspects of authentication systems for voice assistants we identified to address RQ3. These properties showed to be crucial for users' perception of security when performing security-sensitive tasks.

Building Trust

Our participants' perception of security in the context of sensitive tasks on voice assistants was tightly couple with trust in the system. This matches findings about privacy perceptions in shared-user settings by Huang et al. [18] those of bystanders in smart homes by Yao et al. [38]. Participants did not trust a new system out-of-the-box. However, they described several ways to establish trust, especially towards an authentication system. One reoccurring theme was that users transferred trust from a trusted entity to a new system it supports. Interviewees named mostly banks as an example of such an entity, but also *PayPal* and energy providers. Participants stated they would trust a system more if a trusted third party provided it directly. In the words of P8: *"So if it really came from the bank, I'd trust the whole thing more, then I'd be more inclined to use it."* Users apply past experiences to root their trust in entities and are convinced of these entities' interest in keeping their systems secure.

We furthermore found that participants who describe themselves as *"old school"* (P4) were skeptical of novel systems and perceived themselves as less likely to adopt them. Interviewees expected younger users to have an easier time adjusting to a new system. As P10 states: *"It is not normal*

for my mom to do banking on her phone. [. . .] It will perhaps be normal for the next generation to tell Alexa such things."

Positive experiences with a system in the past led to a higher trust in its security. Similarly, users could lose trust by witnessing security incidents. Applying a trial-and-error strategy to authentication can facilitate experiencing a system in a shorter period. Similarly, users could establish trust by checking other users' reviews and ratings. Reading about other people's experiences can have a similar effect on users' trust as experiencing something first-hand. P1 notes: *"If you read that everything works, you have many people who rated this if the reviews are consistently positive, that would certainly build up trust."*

Transparency and Agency

Almost all participants stated that transparency is essential when it comes to the perception of security. A transparent system can enable users to make informed decisions when interacting with such devices. Several participants noted that this property did not transfer well from computers or smartphones to smart home assistants. This attitude was partially due to the fact that voice rendering is difficult to understand for users as the underlying technical fundamentals and information-sharing models are complex.Visual interaction enables users to grasp information much quicker, as stated by P7: *"When I order on the PC, I have several options that I can grasp directly and it is simply easier for me to take in with my eyes than to listen with concentration."* This confirms Abdi et al. [2] who found that visual interaction enables users to absorb information more easily when shopping online.

Using a computer also conveyed a feeling of being in control, which we found is an important characteristic when it comes to security-sensitive tasks. Our findings suggest that using Alexa, in contrast, is perceived as surrendering agency over to another party. Users no longer perceived themselves as the active part and could only hope for the successful execution of the process. They attributed this feeling to an intransparent control flow. P10 states: *"It's weird if I don't see when something happens. Because I say something and it happens. And then I just can't understand whether it was done correctly."*

Our results suggest that the personification of Alexa is a potential factor for this perceived loss of agency. Interviewees compared Alexa to a human operator and expressed that voice commands felt like giving orders to an employee. This perception entailed that Alexa could be affected by human error. P7 explains: *"Suppose I had a butler and I always had to tell the butler: open the front door. I can't trust that* 100% *either. Clearly, somehow, there is a large basic trust. But even then it's kind of uncomfortable when you have in the back of your mind: what if he didn't do it, what if he forgot about it?"* Participants wished for a more transparent control flow, which could lead to an improved understanding of involved entities and task distribution within a system. Voice assistants could accommodate for this by explicitly stating control switches to the back end or third-party services.

Risk Assessment of Authentication

Users' assessment of risks proved to be an important factor in various contexts [18, 21, 30, 39]. Based on their personal assessment, our participants derived variable requirements for an authentication system. We identified an interaction's location as a major factor. Similar to Yao et al. [37], our findings suggest that some locations call for stronger security measures. Most participants agreed that the most distinctive difference was between interactions occurring in a public space (e.g., in front of the door) and those taking place in a private space (e.g., the user's home). Interactions in locations perceived as secure could use weaker authentication mechanisms. P9 states: *"If you're inside the house [. . .] I believe the voice recognition and the code would suffice plenty."*

Some interviewees also distinguished between different zones inside a home. Security-sensitive functionality could be limited to more private areas such as an office or a bedroom. P3 notes: *"Transactions are only allowed from the study, while for the device hanging in the children's room, or in the hallway area where everyone has access, only certain things work there."* Another factor of the risk assessment is being at home vs being away. In accordance with previous work [37], our participants perceived the threat of security breaches to be more prominent while they were away from home. Authentication systems could follow this assessment and apply stronger methods during the vacancy period.

Ruoti et al. [30] highlight how users weigh the perceived risk against the effort needed to protect their privacy. Similarly, we found that participants were comfortable with using weaker authentication mechanisms, if they considered an interaction to be low-risk. Several participants stated that they would prefer having no voice code when checking transactions. In contrast, most participants agreed that an authentication step should be in place to execute transactions. P14 explains: *"I would be fine with using a voice code to see my transaction history, even my account balance, [. . .] but to make a transaction, I don't think Alexa should be allowed to do that."* Some participants also expressed having different requirements of protection depending on the amount of money transferred. Low amounts could be sent without strong authentication. Finally, a few participants explained that, since absolute security did not exist, there has to be a trade-off. *"It just always depends on how much effort I want to put into it, there will be no absolute privacy with such a system."* (P16)

Perception of Authentication Methods

We structured our participants' insights on VUI authentication according to the following four authentication paradigms.

Knowledge-Based Authentication Participants thought of the voice code as a low-level barrier that could primarily mitigate casual attacks and pranks by familiar people. Similar to classic knowledge-based authentication, interviewees were concerned about confusing or forgetting the voice codes for different skills. We found that users could, therefore, revert to code reuse, also across platforms. P9 notes: *"I guarantee you if you have the voice code to open your front door that's gonna be your four-digit PIN for your debit card, it could be for plenty of things in your life."* Similarly, some participants described they would apply coping mechanisms transferred from passwords, e.g., modifying only the last digit between codes. This behavior entails potentially drastic consequences for voice code leaks as they could compromise the security of other systems as well. One participant stated that, while the voice code was not an acceptable authentication method for high-risk tasks, it could serve as duress mitigation. By setting up a code for threatening situations, a user could say that code instead of their usual authentication code, upon which the system would initiate an emergency routine.

Possession-Based Authentication Several participants stated that they would favor token-based authentication with Alexa. Tokens would not be susceptible to the openness of the voice input channel. Hardware tokens could detect the users' physical presence, which should match the Alexa device's location. A close-by token would then lead to the assumption that a legitimate user issued the voice command. P2 gives an example of using a smartphone as a token: *"Alexa can connect to my mobile phone, it's in the same location as I am communicating, then I guess it's fine."* P9 suggested that a microphone-equipped hardware token, such as a *"Fitbit"*, could be used as an authentication token. Devices carried by the user could be marked as trusted, which allows for weaker authentication mechanisms.

Another way how authentication with Alexa could facilitate smartphones would be push-notifications. Some participants expressed that getting a notification requiring confirmation whenever a security-sensitive voice command was executed could be a secure authentication mechanism. Similarly, OTP devices could replace a static voice code. In contrast, some participants stated that using an additional device for authentication would be *"defeating the point of the Alexa, being able to talk to a virtual assistant, now that you have to involve physical things to actually pay, so at that point, you just log into your phone and do it."* (P9)

Biometric Authentication We found that most users preferred biometric authentication due to the natural and effortless interaction with them. However, interviewees expressed concerns regarding the current state of voice recognition on Alexa. P16 states: *"It recognizes you by your voice, but this recognition sometimes doesn't work, and I think that's very rudimentary."* As some participants were aware of possible *Replay Attacks*, they expected future voice biometrics to distinguish live human speech from machine emitted sounds. Interviewees highlighted annoyance caused by false negatives as another drawback of voice recognition. Most participants reported past experiences where voice recognition did not function as expected, possibly due to natural variances in a user's voice. P12 explains: *The voice is often different, let's say when you have a cold, for example. Voice sounds different in the morning than in the evening.* In the context of the scenario *"Door"*, some users also brought up face recognition as a potential authentication mechanism used in combination with a smart home assistant.

Multi-Factor Authentication Some participants proposed combining some of the above-described methods to form stronger multi-factor authentication. Participants perceived that there is a direct relationship between more authentication factors and better security. We found that the preferred combination of authentication methods amongst participants is knowledge-based passcodes with voice biometrics. Other well-known high-risk systems that employ multi-factor authentication, such as bank accounts, probably influenced users' perception.

5 Discussion and Implications for Design

We discuss our main findings (i.e., the themes we identified during the analysis) along with our recommendations for design. We focus on aspects that were perceived as crucial for participants to feel protected during security-sensitive tasks.

Voice Recognition as an Intuitive and Trustworthy Authentication Method

In accordance with previous work [2], our participants found that voice recognition was the most convenient authentication mechanism for voice assistants. It was perceived as a natural way of authentication, as it resembles the human approach to identifying a familiar person, for instance, when talking on the phone. Complementary to known results, we found that this also holds when users perform high-sensitive tasks such as online-banking. Some smart home assistants currently employ a form of voice recognition to distinguish users. However, manufacturers, such as Amazon, do not yet recommend it as an authentication mechanism [5]. Participants were aware of potential shortcomings of voice recognition that researchers and developers need to address before users trust such a system. The most prominently expected feature was liveness detection which distinguishes human voices from speaker playback.

Users want to Test and Experience the Effectiveness of the Authentication Method

We observed that users initially mistrust new authentication mechanisms they had not used before. Some users tried to mimic other users' voice to *test* voice-based authentication. For novel biometric authentication schemes, we recommend including a *demonstration mode* which participants can use to try out the authentication process. Most state-of-the-art systems will block access once a user reaches a threshold of unsuccessful authentication attempts. Such systems are, therefore, not suitable for users to test different adversarial techniques. By including a separate sand-boxed mode that allows unlimited authentication attempts, users might build trust faster and understand novel interaction mechanisms better. Any such demonstration mode must have the same look-and-feel as the standard authentication process, the only difference being that upon successfully authenticating, no real user data is accessible. In this mode, the system should still inform users whether their authentication attempt was successful or not. Reynolds et al. [27] suggested a similar demonstration mode allowing users to verify the functionality of 2FA-tokens immediately after setup.

Users Want Unobtrusive Authentication for Social Situations

We found that participants felt uncomfortable using conspicuous authentication mechanisms in certain social situations. Hence, designs of authentication mechanisms for tasks in social settings need a *discreet mode*. This mode would replace the regular authentication mechanism with an unobtrusive alternative, allowing users to perform security-sensitive tasks without drawing attention to them. While conventional voice recognition has shown to be a desirable option, it might not work for settings that include several bystanders. Situations with considerable background noise make voice recognition inconspicuous, which, however, impedes the correct functioning of the smart home assistant's speech-to-text system, leading to failed authentication attempts. Participants reported having experienced such erroneous behavior before. An implementation of a new system could also automatically identify the current social situation a user is a part of during authentication by, e.g., detecting other persons nearby or measuring the level of background noise. The system could then dynamically adapt the authentication process according to predefined rules for different situations.

Low-Effort Interactions

We identified that effortless and straightforward user interaction are crucial adoption factors. Users reported that their main reason for using a smart home assistant was the low effort interaction with these devices, compared to computers or smartphones. If novel authentication mechanisms diminish the benefit of voice interaction by requiring interaction with other devices, users were no longer willing to use them since the perceived additional risk outweighed the benefits. Also, participants felt that if authentication with a smart home assistant required interaction with a smartphone, they could use the smartphone to perform the task instead. Therefore, the design of new authentication systems for use cases that are already possible with conventional platforms has to consider this risk-benefit analysis made by the users and reduce the effort needed to authenticate to an adequate amount. Such low-effort interaction could be provided by continuous authentication mechanisms, as described, e.g., by Feng et al. [14].

Transparent Authentication Processes

We found that participants were unsure about the information flow of an authentication process. Previous work [2, 38] suggests this is also the case for the general flow of privacy-related data in voice assistant ecosystems. In particular, which party performed the verification of the presented authentication information in scenarios involving third parties (e.g., banks) was not clear to all users. While some believed Amazon would authenticate the user and then get permission to access their account, others perceived Alexa as a literal assistant that takes a user's credentials and uses them to log into an application on the user's behalf. Two factors reinforce the users' perception that Alexa uses third-party systems in the same way a human user would: Alexa's output does not explain whether it came from Amazon or a third party, and users attribute human characteristics to conversational agents. To enhance transparency and make control flow transfers from the Alexa back-end to third-party skills easier to detect, we propose using different voices for each subsystem. This way, a user could instantly notice once the third-party takes over, resolving the aforementioned uncertainty. A similar mechanism to provide transparency could be having Alexa announce handing over control to a skill and reporting back once a request has gone through. This practice could improve users' understanding of the data flow and, consequently, result in more informed security decisions.

Account for Varying Requirements

In line with previous work [21, 38], we found that users have varying security and privacy requirements. In the authentication settings we studied, the two main factors were location and bystanders. In contrast to interactions inside the home, users were concerned about more threats for outside scenarios. In general, users were confident that they could detect malicious behavior from nearby bystanders. Therefore, fewer threats were relevant for such circumstances. Also, the security-sensitivity of the performed task affected the users' security requirements. Most participants agreed that information requests were less security-sensitive compared to tasks

involving money or physical access.

If the principal user was away from home, the smart home assistant should still be accessible or remain turned on. However, security mechanisms should become more restrictive, especially when it comes to authentication. A possible feature accounting for these varying requirements could be a guard mode that, if turned on, requires stronger authentication to turn back off. A real-life example would be an alarm system that only the correct code can disarm. A user could turn on the guard mode if they leave the house or go to bed at night. Upon their return, they authenticate once using a strong and perhaps a multi-factor authentication mechanism to turn guard mode off and switch back to the default authentication method, which could be weaker and less intrusive.

5.1 Limitations

Our sample included almost equal numbers of men and women. We also managed to recruit participants with a variety of educational backgrounds. However, the age distribution of participants skewed towards younger participants. I.e., our study underrepresents older users of smart home assistants. Our sample participants score slightly above average [15] when it comes to the affinity for technology interaction (ATI). CFIP scores indicate that our participants were highly concerned about their information privacy, indicating a further potential under-representation [28]. As this study is exploratory, we targeted users who already have experience with using Alexa. We recruited participants from Central Europe, in part via convenience and snowball sampling. This approach provided us with a potentially limited sample of participants. Hence, our sample might impact how our results generalize to other users. Future work should expand the sample to include different populations, especially underrepresented user groups, such as people with limited visual capabilities or difficulties using conventional keyboards (e.g., upper extremity impairment). Additionally, users of other smart home assistant systems, such as Google Home, might be worthy of further investigation.

We designed our scenarios around a subset of security-sensitive tasks on smart home assistants, namely online banking and smart door locks. As these tasks are currently available in certain markets, we hoped participants might have made some experiences with them. Abdi et al. [3] identified additional security-sensitive tasks, which future work should investigate upon, considering the different circumstances users might experience during the interaction. The most noteworthy tasks, that we did not investigate in our study, are healthcare and home surveillance, as users perceived them as most sensitive.

Some of our scenarios, such as the *"Dinner"* scenario, might not depict real-world use cases that users would want to engage in of their own accord. We deliberately chose edge cases for our study to provoke a stronger reaction from the participants and get richer data. Some of our scenarios include 4-digit PIN authentication, as this is the current standard method on the Alexa platform. However, as other voice assistants include different default settings (see Abdi et al. [2]), future work might benefit from investigating how these different authentication settings impact users' perceptions.

We cope with potential bias introduced by our personal expectations by making our them explicit in Section 3.5.

6 Conclusion

Our interviews explored the design space of authentication for smart home voice assistants. As security-sensitive tasks gain traction on this platform, developers and users call for appropriate authentication measures that enable privacy-preserving functionality and protect data from unauthorized access. Currently used authentication methods such as voice codes and biometric voice recognition proved insufficient considering both casual and targeted attackers. Prior work has already proposed some authentication schemes. However, no previous work has investigated the requirements for authentication systems from a users' perspective.

We closed this gap in the literature by reporting the results of a qualitative user study focusing on security-sensitive tasks on Amazon Alexa. We conducted 16 semi-structured interviews that included four scenarios involving high-risk tasks with Alexa users about (1) their perceptions of threats, (2) mitigation strategies, and (3) design factors that impact secure interaction experience. By performing a thematic analysis, we found that users are primarily concerned about bystanders that can eavesdrop on their interaction with Alexa. Our participants strongly favored biometric voice recognition as they perceived it as a natural and unobtrusive form of authentication. However, most users noted that current systems were not satisfying their security requirements due to being vulnerable to familiar attacks such as the *Replay Attack*.

Based on the insights gained from our user study, we provided design recommendations for future authentication systems. One such recommendation is based on a key finding that users have context-dependent requirements for authentication on smart home assistants. Users perceived levels of risk depending on the location of the interaction (e.g., inside the home vs. outside) and the type of bystanders (e.g., family members vs. casual acquaintances). Participants valued effortless and straightforward interaction with smart home voice assistants. Hence, authentication methods should strictly avoid distracting from primary tasks.

As this study is exploratory, future work can evaluate the findings on a broader basis. Users who have difficulties using traditional computing devices, such as users who can not read well, or users with visual impairment, rely on smart home voice assistants for their daily computing needs. The security and privacy needs and perceptions of this understudied group should be considered in future work.

Acknowledgments

We thank our study participants, as well as our interview partners for the pilot study. Thanks to Simon Anell for helping with transcribing the interviews and Florian Fankhauser for providing feedback on an earlier version of this work. Lastly, we thank the anonymous reviewers and our anonymous shepherd for their valuable and constructive feedback, which was very useful in improving our paper.

References

[1] ATI Scale. https://ati-scale.org/. [Accessed: 2021-02-25].

[2] Noura Abdi, Kopo M. Ramokapane, and Jose M. Such. More than Smart Speakers: Security and Privacy Perceptions of Smart Home Personal Assistants. In *USENIX Symposium on Usable Privacy and Security (SOUPS) 2019*, SOUPS, pages 451–466, Santa Clara, CA, USA, 2019. USENIX Association.

[3] Noura Abdi, Xiao Zhan, Kopo M. Ramokapane, and Jose Such. Privacy Norms for Smart Home Personal Assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI, pages 1–14, Yokohama, Japan, 2021. ACM.

[4] Ruba Abu-Salma, Elissa M. Redmiles, Blase Ur, and Miranda Wei. Exploring User Mental Models of End-to-End Encrypted Communication Tools. In *8th USENIX Workshop on Free and Open Communications on the Internet (FOCI 18)*, FOCI, Baltimore, MD, USA, 2018. USENIX Association.

[5] Amazon. Add Personalization to Your Skill | Alexa Skills Kit. https://developer.amazon.com/de/docs/custom-skills/add-personalization-to-your-skill.html. [Accessed: 2021-02-25].

[6] Simon Anell, Lea Gröber, and Katharina Krombholz. End User and Expert Perceptions of Threats and Potential Countermeasures. In *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, EuroUSEC, Genoa, Italy, September 2020. IEEE.

[7] August. Control Your August Smart Lock with Amazon Alexa | August. https://august.com/pages/alexa. [Accessed: 2021-02-25].

[8] Julia Bernd, Ruba Abu-Salma, and Alisa Frik. Bystanders' Privacy: The Perspectives of Nannies on Smart Home Surveillance. In *10th USENIX Workshop on Free and Open Communications on the Internet (FOCI 20)*, FOCI. USENIX Association, 2020.

[9] Logan Blue, Hadi Abdullah, Luis Vargas, and Patrick Traynor. 2MA: Verifying Voice Commands via Two Microphone Authentication. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, AsiaCCS, pages 89—100, Incheon, Korea, 2018. ACM.

[10] Anthony J. Bradley. Brace Yourself for an Explosion of Virtual Assistants. https://blogs.gartner.com/anthony_bradley/2020/08/10/brace-yourself-for-an-explosion-of-virtual-assistants/, August 2020. [Accessed: 2021-02-25].

[11] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[12] Capital One. Capital One is on Amazon Echo. Questions? Just ask Alexa. https://www.capitalone.com/applications/alexa/. [Accessed: 2021-02-25].

[13] George Chalhoub and Ivan Flechais. "Alexa, Are You Spying on Me?": Exploring the Effect of User Experience on the Security and Privacy of Smart Speaker Users. In *HCI for Cybersecurity, Privacy and Trust*, HCII, pages 305–325, Copenhagen, Denmark, 2020. Springer International Publishing.

[14] Huan Feng, Kassem Fawaz, and Kang G. Shin. Continuous Authentication for Voice Assistants. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, MobiCom, pages 343—355, Snowbird, UT, USA, 2017. ACM.

[15] Thomas Franke, Christiane Attig, and Daniel Wessel. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction*, 35(6):456–467, 2019.

[16] Marian Harbach, Sascha Fahl, and Matthew Smith. Who's Afraid of Which Bad Wolf? A Survey of IT Security Risk Awareness. In *2014 IEEE 27th Computer Security Foundations Symposium*, pages 97–110, Vienna, Austria, 2014. IEEE.

[17] David Harborth and Sebastian Pape. German Translation of the Concerns for Information Privacy (CFIP) Construct. *SSRN Scholarly Paper*, (ID 3112207), 2018.

[18] Yue Huang, Borke Obada-Obieh, and Konstantin (Kosta) Beznosov. Amazon vs. My Brother: How Users of Shared Smart Speakers Perceive and Cope with Privacy Risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI, pages 1–13, Honolulu, HI, USA, 2020. ACM.

[19] Katharina Krombholz, Karoline Busse, Katharina Pfeffer, Matthew Smith, and Emanuel von Zezschwitz. "If HTTPS Were Secure, I Wouldn't Need 2FA" - End User and Administrator Mental Models of HTTPS. In *2019 IEEE Symposium on Security and Privacy (SP)*, SP, pages 246–263, San Francisco, CA, USA., 2019. IEEE.

[20] Il-Youp Kwak, Jun H. Huh, Seung T. Han, Iljoo Kim, and Jiwon Yoon. Voice Presentation Attack Detection through Text-Converted Voice Command Analysis. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI, pages 1–12, Glasgow, UK, 2019. ACM.

[21] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, Are You Listening? Privacy Perceptions, Concerns and Privacy-Seeking Behaviors with Smart Speakers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):102:1–102:31, 2018.

[22] Xinyu Lei, Guan-Hua Tu, Alex X. Liu, Chi-Yu Li, and Tian Xie. The Insecurity of Home Digital Voice Assistants - Vulnerabilities, Attacks and Countermeasures. In *2018 IEEE Conference on Communications and Network Security (CNS)*, CNS, Bejing, China, May 2018. IEEE.

[23] Karen Levy and Bruce Schneier. Privacy threats in intimate relationships. *Journal of Cybersecurity*, 6(1):1–13, 2020.

[24] Nathan Malkin, Joe Deatrick, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. Privacy Attitudes of Smart Speaker Users. *Proceedings on Privacy Enhancing Technologies*, 2019(4):250–271, 2019.

[25] Diogo Marques, Ildar Muslukhov, Tiago Guerreiro, Luís Carriço, and Konstantin Beznosov. Snooping on Mobile Phones: Prevalence and Trends. In *12th Symposium on Usable Privacy and Security (SOUPS 2016)*, SOUPS, pages 159–174, Denver, CO, USA, 2016. USENIX Association.

[26] Dennis Reineck, Volker Lilienthal, Annika Sehl, and Stephan Weichert. Das faktorielle Survey. Methodische Grundsätze, Anwendungen und Perspektiven einer innovativen Methode für die Kommunikationswissenschaft. *M&K Medien & Kommunikationswissenschaft*, 65(1):101–116, 2017.

[27] Joshua Reynolds, Trevor Smith, Ken Reese, Luke Dickinson, Scott Ruoti, and Kent Seamons. A Tale of Two Studies: The Best and Worst of YubiKey Usability. In *2018 IEEE Symposium on Security and Privacy (SP)*, SP, pages 872–888, Oakland, CA, USA, 2018. IEEE.

[28] Ellen A. Rose. An examination of the concern for information privacy in the New Zealand regulatory context. *Information & Management*, 43(3):322–335, 2006.

[29] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit R. Choudhury. Inaudible Voice Commands: The Long-Range Attack and Defense. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18)*, NSDI, pages 547–560, Renton, WA, USA, 2018. USENIX Association.

[30] Scott Ruoti, Tyler Monson, Justin Wu, Daniel Zappala, and Kent Seamons. Weighing Context and Trade-Offs: How Suburban Adults Selected Their Online Security Posture. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*, SOUPS, pages 211–228, Santa Clara, CA, USA, 2017. USENIX Association.

[31] Lea Schönherr, Maximilian Golla, Thorsten Eisenhofer, Jan Wiele, Dorothea Kolossa, and Thorsten Holz. Unacceptable, where is my privacy? Exploring accidental triggers of smart speakers. *arXiv preprint arXiv:2008.00508 [cs.CR]*, 2020.

[32] Jeff H. Smith, Sandra J. Milberg, and Sandra J. Burke. Information Privacy: Measuring Individuals' Concerns About Organizational Practices. *MIS Q.*, 1996.

[33] Elizabeth Stobert and Robert Biddle. The Password Life Cycle: User Behaviour in Managing Passwords. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*, SOUPS, pages 243–255, Menlo Park, CA, USA, 2014. USENIX Association.

[34] Takeshi Sugawara, Benjamin Cyr, Sara Rampazzi, Daniel Genkin, and Kevin Fu. Light commands: Laser-Based audio injection attacks on voice-controllable systems. In *Proceedings of the 29th USENIX Security Symposium*, SEC, pages 2631–2648. USENIX Association, 2020.

[35] Shari Trewin, Cal Swart, Larry Koved, Jacquelyn Martino, Kapil Singh, and Shay Ben-David. Biometric authentication on a mobile device: A study of user effort, error and task disruption. In *Proceedings of the 28th Annual Computer Security Applications Conference*, ACSAC, pages 159–168, Orlando, FL, USA, 2012. ACM.

[36] Rick Wash. Folk models of home computer security. In *Symposium on Usable Privacy and Security (SOUPS)*, SOUPS, pages 1–16, Redmond, WA, USA, 2010. USENIX Association.

[37] Yaxing Yao, Justin Reed Basdeo, Smirity Kaushik, and Yang Wang. Defending My Castle: A Co-Design Study of Privacy Mechanisms for Smart Homes. In *Proceedings of the 2019 CHI Conference on Human Factors*

in Computing Systems, CHI, pages 1–12, Glasgow, UK, 2019. ACM.

[38] Yaxing Yao, Justin Reed Basdeo, Oriana Rosata Mcdonough, and Yang Wang. Privacy Perceptions and Designs of Bystanders in Smart Homes. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):59:1–59:24, 2019.

[39] Eric Zeng, Shrirang Mare, and Franziska Roesner. End User Security and Privacy Concerns with Smart Homes. In *13th Symposium on Usable Privacy and Security (SOUPS 2017)*, SOUPS, pages 65–80, Santa Clara, CA, USA, 2017. USENIX Association.

[40] Eric Zeng and Franziska Rösner. Understanding and Improving Security and Privacy in Multi-User Smart Homes: A Design Exploration and In-Home User Study. In *Proceedings of the 28th USENIX Security Symposium*, SEC, pages 159–176, Santa Clara, CA, USA, 2019. USENIX Association.

[41] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. DolphinAttack: Inaudible Voice Commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS, pages 103—117, Dallas, TX, USA, 2017. ACM.

[42] Linghan Zhang, Sheng Tan, and Jie Yang. Hearing Your Voice is Not Enough: An Articulatory Gesture Based Liveness Detection for Voice Authentication. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS, pages 57—71, Dallas, TX, USA, 2017. ACM.

[43] Verena Zimmermann, Merve Bennighof, Miriam Edel, Oliver Hofmann, Judith Jung, and Melina von Wick. 'Home, smart home' – exploring end users' mental models of smart homes. In *Mensch Und Computer 2018 - Workshopband*, MuC, Dresden, Germany, 2018. Gesellschaft für Informatik e.V.

A Interview Guideline

The guideline we used for our interviews looked as follows, note that italic text indicates actions taken by the interviewer.

Introduction

Greet participant and introduce topic: "Hi, thank you for taking part in this interview." *Present interviewee with consent sheet, explaining purpose of the study.* "In the following, I will ask you some questions where I'm interested in your personal opinions and experiences, so keep in mind there are no wrong answers. If you feel like drawing anything throughout the interview, feel free to use this pen and paper here. Do you have any questions?" *Answer questions of interviewee, if any.* "So let's start with the first question!"

- How long are you using Alexa already?
  - **Alternative:** When was your first contact with Alexa?
- What devices are you using Alexa on?
- Where are those devices usually located?
- What are some typical tasks you perform with Alexa?
- Did you ever use Alexa for online shopping?
  - **If yes:** Did you encounter any issues while doing so?
  - **If no:** Where there specific reasons for you not to use this feature?

Scenarios

Lead over to scenarios: "Thank you for your answers so far. Now I would like you to have a look at some scenarios. For this interview, let's assume that all of the following features are implemented in Alexa, even though some of them are not currently available."

"Now I would like you to please take one of the scenario cards, have a look at it and read it aloud." *Let interviewee choose a card and flip it over.*

For each scenario:

- Please identify any issues that could arise in such a situation?
  - **Follow up:** Why do you think that is problematic?
- Can you identify threats for the user in such a scenario?
- Who could be the source of such a threat?
- What would you do to protect yourself?

Transition to next scenario: "Great, let's continue with the next scenario. However, we can always come back to a previous scenario if you want to add something." *Repeat process for all four scenarios.*

After all four scenarios:

- Now that you have seen all four scenarios, what do you think they have in common?

Demographics

Conclude scenario part and retrieve demographic data: "Thank you for the collaboration so far. Please take the tablet and fill out the questionnaire there." *Hand tablet to interviewee to complete the questionnaire.*

- ATI scale [1]
- CFIP scale [17]
- How old are you? *[free response]*
- What is your gender? *[free response]*
- What is the highest education you have completed? *[Single-select]*
  - *Elementary school*
  - *Junior High school*
  - *High school*
  - *Bachelor's degree or equivalent*
  - *Master's degree or equivalent*
  - *PhD*
  - *Other: [free response]*
- How many people live in your household? *[free response]*
- How many of them use Alexa? *[free response]*

Debriefing

- Do you have any final questions or marks you would like to make?

Thank interviewee for their collaboration and bid farewell: "Thank you again for your participation and have a nice day!"

B Codebook

The following list shows all codes and their categories we used for analysis. The brackets next to the code signify the overall number of occurrences in the interviews.

- **Attackers and Threats**
  - Accidents as threat (49)
  - Amazon listening in on conversations (8)
  - Bystanders as threat (51)
  - Criminals as threat (42)
  - Cyberattacks as threat (40)
  - Insiders as threat (39)
  - Malicious skills as threat (2)
  - Pranks as threat (19)
  - Sharing data with Amazon undesirable (95)
- **Biometric Authentication**
  - Annoyance of false negatives when using biometrics (3)
  - Authentication via voice recognition desirable (50)
  - Risk of false positives when using biometrics (16)
  - Uncertainty about security of voice recognition (16)
  - User wants Alexa in combination with face recognition (17)
  - Voice recognition should distinguish live voice from replays (4)
- **Building Trust**
  - Build/Lose trust through interaction experience (51)
  - Build trust in security mechanism via trial-and-error (6)
  - Trust from reviews (5)
  - Trust in familiar people (41)
  - Trust in system is transferred from trustworthy entity (39)
- **Knowledge-based Authentication**
  - Enter voice code via smartphone rather than Alexa (25)
  - High number of voice codes difficult to remember and distinguish (30)
  - User wishes for duress code (2)
  - Voice code protects against unauthorized access (27)
  - Whispering the voice code protects against eavesdropping (3)
- **Optimistic Authentication**
  - Optimistic authentication does not protect from physical access (1)
  - Optimistic authentication via delayed verification (35)
- **Perceptions of Alexa**
  - Insufficient mental model (17)
  - Personification of Alexa (15)
  - Uncertainty about security of Alexa ecosystem (22)
- **Perceptions of Authentication**

  - Properties of authentication method are transferred from other systems (86)

- **Possessions-based Authentication**
  - Risk of Replay Attacks when using tokens (2)
  - User wishes for Alexa in combination with OTP (25)
  - User wishes for Alexa in combination with token-based authentication (39)

- **Public Sphere of Alexa Interaction**
  - Openness of voice interaction security/privacy relevant (85)
  - Reconnaissance of Alexa easily possible (4)

- **Requirements of Authentication**
  - Multiple users use Alexa in parallel (10)

- **Risk Assessment of Alexa Authentication**
  - Minimal protection by law (8)
  - Users notice acoustic attacks on their Alexa if they are present (3)
  - User wishes for multiple authentication steps (37)
  - Variable security requirements depending on location (42)
  - Variable security requirements depending on presence of user (9)
  - Weighing up risks and effort of authentication (65)

- **Risk-Benefit Analysis of Alexa**
  - Alexa needs justification to exist (117)
  - Refrain from using the system due to security reasons (37)
  - Weighing up use against increased exposure to risk (30)

- **Social Aspects of Alexa Use**
  - Hierarchy among Alexa users (1)
  - Take time for important actions (6)
  - Using Alexa means being lazy (24)
  - Voice interaction inappropriate in specific social situations (31)

- **Transparency and Agency**
  - User wishes for agency over transparent processes (86)

- **Users' Mitigation Strategies**
  - Build trust in security mechanism via trial-and-error (6)
  - Change voice code regularly (11)
  - Move to another room to use Alexa (30)
  - Refrain from using the system due to security reasons (37)
  - Take time for important actions (6)
  - Users notice acoustic attacks on their Alexa if they are present (3)
  - Voice code protects against unauthorized access (27)
  - Voice interaction inappropriate in specific social situations (31)
  - Whispering the voice code protects against eavesdropping (3)

C Scenarios

Figure 1 shows the scenarios used in all our semi-structured interviews. We printed these for in-person meetings, we showed them on a website for online meetings, and made an audio version that included image descriptions for a blind participant.

You gathered some friends for a dinner party at your place. In the middle of eating you remember that you owe Kim 20€ for the lunch she paid the other day. You want to settle this right away.
You say: *„Alexa, transfer 20€ to Kim!"*
Alexa responds with: *„OK, to transfer money, tell me your voice code!"*
You: *„My code is 8915."*
Alexa accepts the code and the transaction succeeds.

(a) Dinner

You are in your living room watching TV when your partner asks, if you have already paid the utility bill this month. Since you have in fact not done so yet, you decided to do it right away using your Alexa device.
You say: *"Alexa, pay the utility bill!"*
Alexa answers: *"OK, to pay it, tell me your code!"*
You: *"6858"*
Alexa accepts the code and the payment is processed.

(b) TV

You have just taken all your groceries out of the car and are about to take them inside. The front door is locked. Your hands are full and you don't want to put everything down again so you ask Alexa to do open it for you.
You say: *"Alexa, unlock the front door!"*
Alexa answers: *"OK, to unlock the door, tell me your voice code!"*
You: *"3071"*
Alexa confirms the code and the door is unlocked.

(c) Door

You just came back from working in the garden. Your kids run around the house screaming. They are already very excited for the upcoming school trip. That's when the question comes to your mind: have you already paid for that? You want to check if the transaction is there in your online-banking.

(d) Hands

Figure 1: Scenarios used in the semi-structured interview

"The Thing Doesn't Have a Name": Learning from Emergent Real-World Interventions in Smart Home Security

Brennen Bouwmeester, Elsa Rodríguez, Carlos Gañán, Michel van Eeten, Simon Parkin
Department of Multi-Actor Systems (MAS), Delft University of Technology
b.j.bouwmeester@student.tudelft.nl
{e.r.turciosrodriguez, c.hernandezganan, m.j.g.vaneeten, s.e.parkin}@tudelft.nl

Abstract

Many consumer Internet-of-Things (IoT) devices are, and will remain, subject to compromise, often without the owner's knowledge. Internet Service Providers (ISPs) are among the actors best-placed to coordinate the remediation of these problems. They receive infection data and can notify customers of recommended remediation actions. There is insufficient understanding of what happens in peoples' homes and businesses during attempts to remediate infected IoT devices. We coordinate with an ISP and conduct remote think-aloud observations with 17 customers who have an infected device, capturing their initial efforts to follow best-practice remediation steps. We identify real, personal consequences from wide-scale interventions which lack situated guidance for applying advice. Combining observations and thematic analysis, we synthesize the personal stories of the successes and struggles of these customers. Most participants think they were able to pinpoint the infected device; however, there were common issues such as not knowing how to comply with the recommended actions, remediations regarded as requiring excessive effort, a lack of feedback on success, and a perceived lack of support from device manufacturers. Only 4 of 17 participants were able to successfully complete all remediation steps. We provide recommendations relevant to various stakeholders, to focus where emergent interventions can be improved.

1 Introduction

The use of "smart" Internet-of-Things (IoT) home devices amongst consumers is growing, where this can include internet-connected home appliances, entertainment systems, and home fittings such as smart doorbells or locks. The connectivity of these devices has historically lacked sufficient security [1, 23]. Many commonly-used IoT devices have not only technical vulnerabilities, but also ineffective configuration options for password and access permissions [3, 17]. This means that a range of consumer IoT devices continue to be susceptible to malware infections, facilitating various forms of abuse, from recruiting them into botnets to personal stalking and harassment [51].

There is a direction of travel to ensure that consumers purchase secure devices, e.g., increased awareness [48], labels indicating security properties [22,47], and improved standards of device design [11]. However, for the foreseeable future, insufficiently secure devices continue to enter the consumer market. The brunt of the efforts to clean up infected IoT falls on both the end-users who own the devices and Internet Service Providers (ISPs), where more than 80% of the devices are located [14].

RFC6561 states that ISPs should notify users and ask them to remediate the threat [44]. Helping users protect their computer systems and remove infections has proven to be difficult for PC-based malware, even where users are more likely to have workable, effective tools available to them (for instance, automatic OS update mechanisms [74]). In the consumer IoT space, the conditions for user advice and remediation can be much more constrained when it is an ISP contacting a customer with advice; it is usually unclear what exact device, or even general device type, has been infected, forcing the advice to be highly generic. The lack of accessible user interfaces makes it difficult for users to perform the required security actions on the device they suspect is infected.

Prior work has found that notifying a user about an IoT infection can lead to cleanup [14]. Much less is known about the processes which take place in end-users' homes after receiving a message with remediation advice. When technical experts are approached to clean a 'smart' personal device of suspected malware or unwanted code, they may not be able to confirm it is infected or prove removal of malware [33].

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

We conduct our study by partnering with an ISP which has sent notifications with remediation advice to customers infected with Mirai malware. We specifically report on the experiences of 17 ISP customers in their efforts to apply the advice. Mirai is a malware family that came to prominence in late 2016 [6], and has been referred to as the "king of IoT malware" [49]. It continues to be the leading malware family [39]. Following the notifications, we approached customers to conduct remote think-aloud observations of their attempts to follow the advice in their home, surrounded by a variety of potentially affected devices.

We focus on the following question: *How do end-users act on remediation advice about their infected Internet of Things device(s)?* To answer this question, we documented the end-to-end story of botnet remediation which included network measurements to identify affected users, and device owner engagement. Infection data received by the ISP allows us to identify users with an infection, but also to gauge the remediation success after the intervention. We combine this with qualitative data collected during the think-aloud observations. We make the following contributions:

- We report on the real-world, in situ experiences of 17 customers acting on advice for IoT devices suspected to be infected with malware. We step out of controlled lab conditions where advice that has a known outcome is directly provided to participants. This allows us to collect data with higher ecological validity.
- We show that users are motivated, yet the advice is constrained by what can be known about the location of the infection on a home network. Many recommended actions are in practice outcomes which users must find a way to reach based on behaviours familiar to them. This adds detail to the shortfalls in the last part of advice communication for smart home users – the implications of the best-placed stakeholders (the ISP) intervening to communicate advice which is the best-available practice or which has been consolidated from manufacturers, to context-expert end-users.
- We capture the importance of advice signal design for effective behaviour change relating to smart home security hygiene. For this we relate our results to the Fogg Behaviour Model [27]. We find that where the Activation Threshold for supporting an individual to reach a target behaviour is often treated as if it were a line to cross, with home IoT it is more akin to an 'Action Diffraction'. The user is not *able to do enough* in a direct path to the goal, due to limitations inherent in the environment, such that advocated best-practice behaviours are non-deterministic. Participants applied a range of behaviours in an approach that appeared to have a good chance of working but which were not definitely going to be successful, or be confirmed as having been successful.

The context of malware infections of consumer IoT devices is discussed in the Background (Section 2), including how users are typically engaged to remedy consumer IoT infections. We describe our Methodology in Section 3, and Results from our in situ sessions with participants in Section 4. The implications of our participants' experiences are discussed in Section 5 and contrasted with Related Work in Section 6. Concluding remarks and directions for future work close the paper in Section 7.

2 Background

Many devices enter the market that lack even basic security precautions [3]. The existence of a botnet such as Mirai starts with the manufacturing of IoT devices, which are then shipped, bought by retailers and later by consumers. Once a device has been infected, it is also unclear which of these stakeholders carries the responsibility for cleaning the device, but manufacturers generally lack incentives to prevent and remediate this problem [65].

2.1 Attacks on consumer IoT devices

Different malware families use different vectors to infect vulnerable devices (such as routers, cameras and digital video recorders) [6, 14]. In the case of Mirai, there are four stages [6, 19, 38, 45, 68]. The first stage is to perform a brute-force attempt to access the device using a sequence of entries from a list of standard known username/password combinations. If this brute-force succeeds, the newly infected device sends its IP and username/password combination to the attacker. In the third stage, the report server informs the loader, which loads the malware binaries onto the device. After the binaries have been executed successfully, they are deleted, and the device is now part of the botnet.

Many IoT devices do not support standard user interfaces, which makes it difficult for customers to change the standard passwords (assuming a device has such a feature to begin with, which may not be the case [26]). Even where a device has an adequate interface, many users prefer having a working device as soon as possible over going through security-related installation steps (such as replacing the standard password) thoroughly [40] (where the inter-connected nature of smart homes means this may include securing the entire home network). End-users who do care about security may lack knowledge to perform the right actions, due to the heterogeneity of IoT devices [5, 78].

2.2 Improving consumer IoT security

Information about the security qualities of IoT devices can potentially be difficult to find. One avenue of research focuses on supporting consumers to make informed choices about the smart home devices they buy in the first instance (e.g., security

labels [22, 47] and consumer guides [48]) Another area of focus has been to ensure that device design matches user needs; this has been noted regarding specific requirements for access control [34] and privacy in a shared environment [77], for instance.

Most vendors of IoT devices do not deliver a comprehensive manual or support page with their product. Where information is provided, details relating to security are often absent or not adequate [8, 30]. This means that even for those consumers who do care about security [9, 50, 64], the 'transaction costs' of ensuring purchase of the most secure device are simply too high [2, 8].

As the Internet increasingly connects end-users and their devices globally, it becomes complex for governments across the world to organise clear responsibilities and liabilities for security. As the IoT is still relatively new and evolving, it could take some time before governments are able to clean the market of insufficiently secure devices and exert pressure on responsible parties. Simple improvements such as labelling the level of security of devices could improve the purchasing environment [37], but even for such small improvements, incentives are lacking. As present, the most viable mitigation techniques mostly come from Internet Service Providers (ISPs) intervening when customers' devices are compromised, or information campaigns to realise prevention through consumer awareness. However, levels of remediation are far from perfect. The content of a notification should be understandable and clear for target users, but there is a balance to be struck. Research has found that detailed steps can strengthen the effect of the notification [21, 43, 72]. On the other hand, messages should be plain and simple [29].

Even where users are aware of a security problem and activated to act, there can be uncertainty about which device is infected, or how to take the required action [60]. Users may instead rely on familiar techniques to solve problems on 'unfamiliar' devices, which often is not the correct approach for new types of devices and infections [76].

For structuring interventions, identifying critical points in life cycle of devices is useful [41]. *Opportune moments* for intervention then emerge [27], which are important for focusing resources toward enacting a behaviour at a specific point where it is more viable. Where purchase of new devices is one such point [22, 54], the notification to a customer of a suspected malware infection is another opportune moment. However, There are challenges inherent to deploying behaviour interventions where the 'influencer' does not manage the environment. In managed environments (including the artificial/controlled environment of a lab study), the influencer can know who the target is and how to reach them. Here, we study an environment where that knowledge is not immediately available. We then leverage technical tools to approximate where the intervention is needed, by triangulating across datasets to identify devices which are vulnerable. Simply put, we have to find a way to go to the participant, whereas normally in a study the participant comes to us.

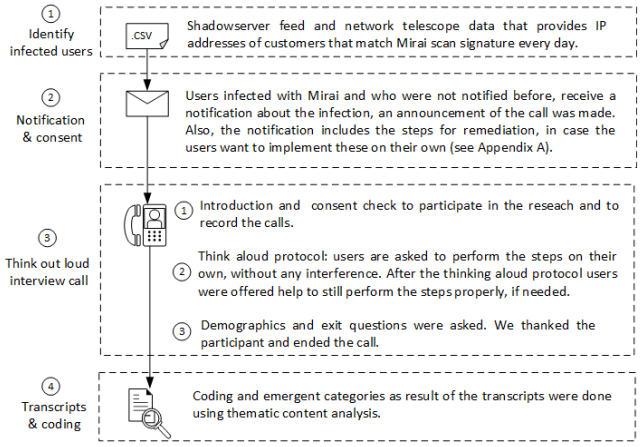


Figure 1: Approach and data collection.

3 Methodology

In this section we describe our approach to answering the main research question. This involved partnering with an Internet Service Provider (ISP) and studying customer responses to remediation instructions.

3.1 Overall approach

Our study starts with identifying ISP customers who suffer from an active Mirai malware infection. For this, we used two data sources. One was the Shadowserver drone report [67]. The ISP receives from Shadowserver a daily list of IP addresses of customers that match the Mirai fingerprint. Mirai scans have a particular signature, where an artefact of the malware's stateless scanning approach is that each probe includes a TCP sequence number equal to the destination IP address that the malware is targeting to attack [6]. This is conventionally used to detect the malware.

A network telescope was then employed. This is a set of unused IP addresses [46], where the traffic targeting this IP set is usually unsolicited. The network telescope of 300K IP addresses logs the IP addresses of hosts that were scanning with the Mirai fingerprint, as described in [6].

This is Phase 1 in the overall approach (as in Figure 1). The ISP is in a unique position to know which customer is associated with an IP address, so that we could identify which customers were suffering from a Mirai infection.

If the identified owner had not yet been notified, the ISP would notify the user about the infection via email (Phase 2, Figure 1). Included in this email would be an explanation of the research, and an invitation to participate in a call to understand better the process that users follow to execute the steps, as part of the standard service. It is also mentioned

that users are free to execute the steps themselves (see Appendix A for more details on the notification) without opting in to the study. During the call, each customer was asked explicit consent to participate in the research and record the call (see Appendix B). Minimal data of customers who did not consent to be part of this research was received in advance to be able to contact the customer, but it was not included in the results of this research.

To further ensure that the email notification could be understood by those end-users who received it, several communication experts from the communication department of the ISP transcribed the text to B1 level of the Common European Framework of Reference of Languages (CEFR) [25]. This is an international standard to describe language proficiency, in which B1 indicates basic level. The email notification was written in both English and Dutch (as the main language where the study was carried out).

A day after the email notification, users would be called (Phase 3, Figure 1). Three users did not answer during three attempts to call them and were left out of the study. Our protocol has a check at the beginning to ensure we talk to the device owner. We then asked users whether they wanted to opt into participating in the study, asked for explicit consent to record the interview, and explained that the participant could end the call at any moment (Appendix B, part 1).

After concluding a call, a transcript was created. We used thematic analysis (Phase 4, Figure 1) to code transcribed copies of the interactions (from audio recordings). For performing the thematic analysis, the step-wise approach listed by [4] is used. Two of the researchers coded the transcripts to identify themes. Dedicated code review discussions took place between coders (to address emergent themes and conflicts), which happened in stages before arriving at the final set of themes. A balance in themes was found through iterative merging and splitting existing themes until convergence was reached into the most important themes (where the subsection in our Results represent theme families, Section 4). Saturation of themes was reached after 17 calls.

3.2 Think-aloud protocol

Originally we had planned to visit customers' homes/premises, to interact with them in an a natural and comfortable environment, and be physically present when users execute the recommended remediation advice. There was a need to instead develop a novel phone-based protocol for interacting with the customers of the partner ISP, foremost due to social distancing measures (Section 3.6). A positive aspect of this was that all participants were at the appropriate location when they were contacted.

To prepare, experience was gained in managing cases where remediation was not possible. One of the researchers accompanied a senior mechanic from the ISP for a day, and gained insights from the ISP customer support staff regarding how to build trust with customers. In cases where the engineer is not successful in helping users, the most important step was seen as informing the consumer of the situation and to let them know about the possible ways forward. In such cases, also a supervisor should be informed about the issue. It can reach a point where informing the customer of an issue is the best one can do. This reflects the reality that the ISP is not technically responsible for the device, even though it has the opportunity to intervene.

The think-aloud protocol (Phase 3, Figure 1) consisted of three stages:

- **Stage 1: Consent and notification:** First, we obtained consent to conduct the study, asking then for approval to record the interview. Next, we checked whether participants received the notification and, if not, we sent it again and provided the participant time to read it.
- **Stage 2: Acting on the advice:** We allowed the participants opportunity to perform the actions and verbalize their thoughts, without direct input from the researcher. This think-aloud activity was transcribed and analysed.
- **Stage 3: Demographics and support:** We collected demographics and, if the researcher saw an action during Stage 2 as incomplete or incorrect, suggestions were offered for performing actions correctly, to the extent that this was possible (see 3.7). Last, we thanked the customer for their participation as well as provide e-mail details for future contact with the researcher in case they had any questions.

See Appendix A for complete details on the think-aloud protocol. The technical advice provided to customers (in the email and in the second step of the protocol) are steps used by the partner ISP, so it is what the ISP considered best advice. For comparison/reference, these steps are comparable to what is advised in online sources, as found on the Krebs on Security blog[1] and Symantec/Norton website[2].

During a call with a participant, they would try to implement the 5 recommended actions from the email: (1) determine which devices are connected to the internet that could potentially be infected with Mirai; (2) change the password of these devices; (3) restart the devices by turning them off and on; (4) reset the modem/router to the factory settings, and; (5) change the password of the modem/router (Appendix A contains the message in full).

3.3 Pilot

The study protocol was tested with 7 customers. These pilot sessions were especially important for refining the protocol,

[1] https://krebsonsecurity.com/2018/01/some-basic-rules-for-securing-your-iot-stuff/

[2] https://us.norton.com/internetsecurity-iot-smart-home-security-core.html

Table 1: Summary of participants demographics, devices, actions, and outcomes. No. of users refers to the number of people in the household of the participant. Some connections were part of a small business rather than a home. Steps 1-5 refer respectively to actions relating to Device Identification, Device Password, Device Reset, Router Reset, and Router Password. Boxes highlighted in gray refer to an outcome classed as a failure to complete the associated Step, otherwise the action was a variation on a successful outcome. The letter-specific codes for each step are detailed in Figure 2.

Index	Age	Gender	No. Users	Suspected device	Step 1	Step 2	Step 3	Step 4	Step 5	Remediated?	Reinfection?
1	53	M	6	Router	1B	n.a.	n.a.	4A	5A	Yes	No
2	55	F	1	IP camera	1A	2D	3A	4C	5C	Yes	No
3	43	M	2	IP camera	1A	2D	3A	4C	5A	Yes	No
4	49	M	3	IP camera	1A	2D	3A	4C	5A	No	Yes
5	65	M	2	IP camera	1A	2C	3A	4D	5D	Yes	No
6	21	M	Business	IP camera	1A	2B	3C	4C	5A	Yes	No
7	45	M	4	Router	1B	n.a.	n.a.	4C	5C	Yes	No
8	65	M	2	NAS	1A	2C	3A	4C	5A	No	Yes
9	61	M	2	Smart printer	1A	2C	3A	4C	5A	Yes	Yes
10	34	M	Business	IP camera	1A	2A	3B	4A	5A	Yes	No
11	55	M	Business	NAS	1A	2A	3A	4A	5A	Yes	No
12	80	M	2	Doorbell	1A	2A	3A	4C	5A	Yes	Yes
13	49	M	1	IP camera	1A	2D	3A	4A	5A	Yes	No
14	43	M	2	-	1C	2E	3D	4A	5A	Yes	Yes
15	53	M	5	Router	1B	n.a.	n.a.	4B	5B	Yes	No
16	41	M	3	IP camera	1A	2B	3C	4C	5A	Yes	No
17	42	M	4	Smart TV	1A	2C	3A	4A	5A	No	No

as the main study would also involve interacting with real customers of the ISP and an intervention that has not been studied directly in a real-world setting. We could also evaluate the think-aloud protocol, accounting for not being present in the room with the users.

Similar to the insights from the ISP customer support staff, trust was found to be important: 5 of 7 customers were cautious about the call, 4 wanted a more detailed explanation of the research, and one called back to the service desk to confirm the authenticity of the research and email.

The pilot resulted in a check being added at the beginning of the protocol to talk to the person who takes care of security issues (as pilots included cases where the person who set up the devices did note live in the household); issues of delegation to informal technical support are discussed in [56]. The most significant change in the protocol was the inclusion of more upfront information about the purpose of both the call and research, to bolster trust.

3.4 Participants

All customers with a diagnosed Mirai infection in the period between May and July 2020 were notified by email about the infection and the study. If they did not opt out of the ISP's support process, they were called the next day. During the experiment period, 37 unique IP addresses corresponded to 37 customers with Mirai infections. 12 were observed during the weekend, where the helpdesk at the ISP does not notify these users as they cannot provide support over the weekend. Of the 25 remaining IP addresses, 3 could not be notified due to technical issues within the ISP, 2 did not respond to attempts to contact them after being notified, and 3 were not willing to take part in the experiment (did not opt-in to the study). There were think-aloud observations with 17 customers. The age of the participants was between 21 and 80 years old with a median age of 49. We interviewed 16 males and 1 female, and from the 17 participants, 3 used their internet connection to run their own businesses. Table 1 shows the participants' demographics. As was also the case during the study pilot, sessions each took approximately 30 minutes in total (15 minutes of which was the think-aloud protocol).

No incentive was provided to users to participate, beyond the possibility of providing the technical support detailed in the participant-facing study materials (see Appendix).

3.5 Measuring cleanup

From the two data sources described in subsection 3.1, we received daily lists of IP addresses where infected Mirai hosts were located. This led to the initial identification of the customers and the recruitment of participants. We kept monitoring this data for an additional two weeks after the call.

Mirai reinfection can occur within a few minutes, or for some devices within 48 hours [14]. We chose a conservative 4-day window to determine remediation. Since Mirai attacks involve aggressively scanning the IP space for devices, we presumed a two-week window to measure reinfections as related to the state of participants' home network. We illustrate this way of measuring outcomes in Table 1. We should note that this observation method is not perfect. While false positives

are highly unlikely, because of the specific Mirai fingerprint, false negatives might occur (an infected host might not show up in the data, even though it is still infected).

3.6 Ethics

The study protocol was approved by our institution's human research ethics committee (TPM project 1083). The study design followed the principles for ICT human research as detailed in the Menlo Report [20] (as indicated also in the design of the think-aloud protocol). To make sure the end-users feel that they are in a safe environment, the think-aloud protocol is built around ensuring that the participant feels they are in a safe space and have not done anything wrong, and can state their feelings and actions without any judgement.

The first part of the call is about informed consent. This consent involves both taking part in this research anonymously, as well as the call taking place and the recording of it. Users were reminded that they could stop the study at any time. If they did not wish to participate, they were informed that they would be processed as usual by the partner ISP.

3.7 Limitations

In adherence with national social distancing measures related to the Covid-19 pandemic, in-person data collection was avoided. In-person home visits may have allowed for opportunistic observation of relevant details outside of our protocol, or differences between stated and actual behaviour. We compensated for this with a think-aloud protocol. We cannot rule out, however, that users may not have accurately described what they did via the call. Even though the researcher is trying to stay at the side-line, their presence influences the participants [36, 42, 71], who will typically pay more attention and effort to the tasks within the study. This does not detract from the context of the interaction, which would naturally require the individual to focus on the instructions regardless.

The research may have engaged with device owners who were unable to knowingly secure their devices. In such cases, at the end of the protocol they were helped to execute the steps they missed properly (after the think-aloud protocol). Also, an e-mail for future questions or contact was provided. The researcher helped the participants with any unsuccessful steps in accordance with the study protocol. Although infections could have plausibly been remediated, participants were carrying out actions themselves within the online 'interview call' format, and outcomes were based on customers' reported actions. For instance, users may have changed passwords though we may not have been able to corroborate the outcome, or whether the advice absolutely caused the outcome.

Our work is based on users' data from a single ISP. Hence, more research will be necessary to validate these results across multiple ISPs and different countries. Similarly, we focus our design and analysis on a single malware family, Mirai. The recommended steps might differ from those for other malware families. We see trends of advice only becoming more complicated, see Section 5.2).

A final point is that our measurements of remediation and reinfection is not perfect. The infection data suffers from a small rate of false negatives. We compensate for this by working with longer time windows. Only when participant's IP addresses are not seen in the infection data for four consecutive days, do we conclude they successfully remediated.

4 Results

Participant sessions were transcribed and analyzed to understand the 'journey' of remediation, following the steps of advice. We present our findings by following this journey. No participants reported having attempted to apply the steps before the session. We describe how participants attempted to: first, identify the infected device (Step 1, subsection 4.1); implement the recommended actions on that device and on their router (Step 2-5, subsection 4.2); infer the success of their actions (subsection 4.3), including their motivation to work through what transpired to be an arduous process for almost all participants (subsection 4.4). Finally, we connect the customer experiences with our measurement data on whether the infection was remediated (subsection 4.5).

Figure 2 provides an overview of reported participant actions. Each labelled box represents a particular action. To illustrate: 13 users took action 1A and identified a specific device as infected. White boxes indicate a successful action in terms of enacting advice, grey indicates no success.

4.1 Identifying suspect devices in the home

The first remediation action is to identify which devices are connected to the internet and could be infected with Mirai. The notification email informed participants that Mirai would not be present on a regular PC, laptop, tablet or phone. The subsequent actions (changing the password and turning the device off and on) are meant to be applied to all the devices that could potentially contain Mirai. A cautious approach is then to remediate and secure all potential victim devices.

Thinking aloud, four participants immediately focused on the device that they thought was the most likely culprit. All other participants started enumerating their devices, e.g., P12: "*I have 22 devices connected to the internet. Cameras, a garden sprinkler, a doorbell, the list goes on.*"

Whether multiple devices were enumerated or not, all participants focused on identifying one suspect – no participant ended up identifying multiple suspect devices. We observed participants using three heuristics to reason about the likely culprit. The first heuristic, used by the majority of participants, was a process of elimination, as with P04: "*I have a laptop, two mobile phones, no three mobile phones. I have a camera, a security cam, and the solar energy is also connected to the*

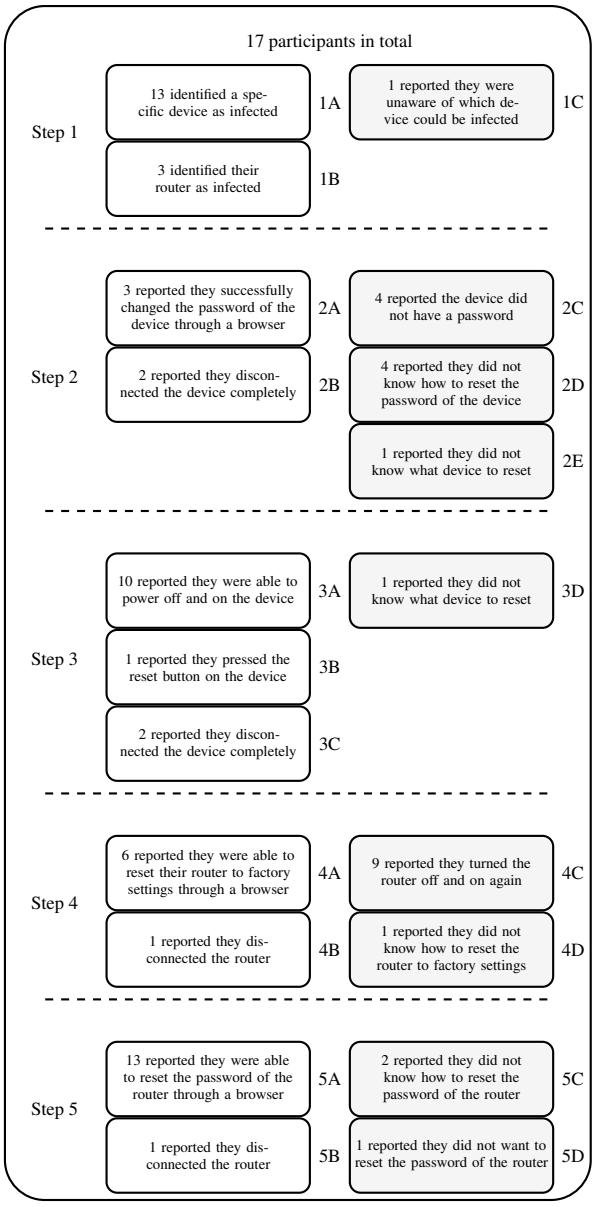


Figure 2: Overview of outcomes of actions by participants, while attempting to execute the remediation advice. Steps correspond to those found in Appendix B.

internet. I run anti-virus on everything. I just bought that for five devices, also for my wife's iPad. According to that email, it would have to be the security camera."

This first heuristic might not lead to a confident identification, as seen with P01: "*OK, in the email you write that it can't be phones, laptops, or really anything with Android on it. That leaves us with printers and cameras and the like. But I don't have those. Yeah, I have a printer, one of those all-in-one types, but that isn't even switched on at the moment [...] So that doesn't make sense.*"

The second heuristic, used by eight participants, was honing in on a device that the person recently experienced problems with. This occurred for instance with P02: "*I think it is the camera. [...] It says there is a system error and it needs a restart. But only the company can do this remotely.*", and P06: "*There are 4 phones connected to the wifi and a computer. And the security camera, but that doesn't work properly anymore. It actually seems likely that this camera is misbehaving.*"

A third heuristic was only employed by one person: conducting an Internet search. P15: "*I have one all-in-one printer, that is never turned on, a beamer connected to the internet, an Xbox, Nintendo Switch, a smart TV, 2 laptops with Windows 10, a laptop with Windows 8 and a [routerModel] [...] Now, I saw in the email that it can't really be one of these devices, so I searched on Google for all my devices [...] then I found that [routerModel] has been having problems in [another country], so that was really the only clue I could find.*"

In one case, the participant enumerated the devices they owned, but felt uncertainty around finding the offending device made the whole process meaningless. It is interesting to note that all participants experienced this kind of uncertainty, but only P14, who indicated they had technical expertise, felt it invalidated the remediation path: "*Can you see something useful, like an IP or MAC address or something? [...] I have no idea [what device could be the problem], so half of these steps I can't execute. That makes this process kind of useless.*"

4.2 Taking action with a suspect device

Only three participants reported that they were able to change the password of the suspect device (Fig 2). In these cases, the device either had an associated app or an interface on the device itself that allowed the user to initiate the password change. For, P11, who owned a Network Accessible Storage device (NAS): "*Yeah, resetting the password, you can do that via a small screen [...]. It worked, now with a slightly more difficult password.*"

Four other participants indicated that they thought the device did not have a password, e.g., P09: "*This [printer] has no password, does it? I can search on the internet, but I think the printer just appears on screen when I want to print. Other than that, there isn't much to it. I don't get any hits when I search for something related to passwords.*"

Four participants said they did not know how to change the password, as with P03: "*Well, I really have no idea how to do this. I do not have a booklet or anything. And the thing has no name, I think. So you tell me how to do this. A friend of mine helped me with installing this thing, but he got killed in a car accident, so I can't ask him.*" One participant consulted the manual, P17: "*There is really nothing useful in the booklet that comes with it. I only see things that prevent us from suing them.*" Two participants reported visiting the manufacturer website, to no avail, as for P13: "*Yeah, I searched for this and I found a website that belongs to the device. But the site is totally unhelpful. I already know it is a camera, can't they put*

something more useful on the site?"

Two participants 'solved' the problem by completely disconnecting the device, e.g., P06: "*You know what, I will just disconnect it. I have no idea how to change the password, but it is broken anyway, so I will take it offline and then we will buy a better one [. . .] I don't want a virus in my network.*". P16 followed a similar behaviour: "*Well, I thought that [the camera] would hang there as a deterrent. But then I got your email. I threw out the device right away, because I definitely do not want a virus.*" Chalhoub & Flechais [15] considered disconnecting a smart device as a *compensatory behaviour* that owners apply to address security and privacy concerns, regardless of whether it directly addressed the concern.

When it comes to restarting the suspected device, two participants looked for a dedicated reset button. P10: "*I am pressing the reset button for a long time [. . .] OK, it is turning off and on again.*" The second person looked for such a button but ended up, like nine other participants, disconnecting the power cable: P02: "*I don't really see a button or anything on the camera. Perhaps just pulling the plug then?*"

The last two steps concerned the modem/router. At least six participants had the standard router issued by the ISP. The email from the ISP contained a link to a help page that described two actions: how to restart the device by disconnecting the power, and how to factory reset the device via a web interface. While the email asked users to factory-reset the router, the presence of both actions on the help page led some participants to take the first listed action: only disconnecting the power. Strengthened by the presence of this action on the help page, participants were convinced their efforts were the requested ones, P02: "*It says here to pull the plug and wait for 10 seconds, I can do that, great*". Moreover, participants tend to copy the actions they took for earlier steps and implement those for their router, P08: "*Reset? So I will do the same as with the camera. I have disconnected it for 5 seconds and it is back in. I see a green light so I guess that worked*". Overall, 6 participants reported having enacted a factory reset, while 9 participants removed the power cable to reset the device.

P05, who was running a small business, said they did not want to execute a factory reset: "*The problem is that I would have to set up all port forwarding again and I don't really want to do that [...] Then I have to let IT come again. [...] Were the previous steps not enough to make the virus disappear?*"

For the final step, 13 participants reported that they successfully set a new password via the ISP web interface of the device, while two said they did not know how to do this. For this step, six participants made use of the URL in the notification (see Appendix A).

4.3 Inferring the success of remediation

When users manage to complete an action on the suspect device, they receive almost no feedback on the success of their efforts. The exception was when setting a new password was supported via an interface that the participants are familiar with. The users who managed to reach a web interface for their router, for example, would get a clear confirmation when they successfully completed a password change. Still, all participants experienced actions that lacked feedback on whether they were successfully completed. More importantly, all participants lacked feedback on the success of their actions in terms of the main outcome: removal of the malware. These observations are of interest when compared to Forget et al. [29], and the examination of whether 'engaged' or 'disengaged' users arrive at secure outcomes to their (in)action to secure a computer – here the problem is that the outcome, secure or not, is not visible.

During the calls, we witnessed a clear desire by many participants to receive confirmation of whether they were doing the right things, as with P02: "*Shall I wait a few seconds? [...] OK, I think 10 seconds is enough, I am putting the plug back in [...] I am waiting for the lights to turn on again. It is supposed to be orange, right? Or green?*"

Some remediation actions were surrounded by uncertainty, while others were more clearly unsuccessful to the participants. In either case, participants regularly requested confirmation that they were successfully removing the virus. For instance, P04: "*Could it be enough if we do not change the password. That we do all other steps?*", and P08: "*The device is already disconnected. Does that count as a reset if I now reconnect it again? I am really curious whether the virus is really gone. Can I reconnect it now?*".

4.4 Motivation under uncertainty

All participants were willing, in some cases eager, to undertake the recommended actions, e.g., P09: '*'I am now putting the plug of the router back in. What is the next step of this adventure?*" Participant motivation was illustrated by the degree to which they tolerated their uncertainty about what was asked of them, and whether they conducted the actions correctly. Motivation was also visible through the effort that was made. For example, the device or router might be in another part of the house or access to it might be blocked. This was the case for P03: "*You ask quite a bit from me, because then I have to make quite a mess. [...] Let me put the phone down, I need to move a few boxes... OK. What do I do now*?", and P07: "*Then I will walk to the utility closet [...] I see the cable already, I will pull it out completely.*"

In addition, the factory reset of the router means that users lose their configuration, which might not be trivial to set up again. P10 debated this, "*Ah, so then I have to set up all port-forwarding and port assignments again. Well, I think that is the right thing to do, otherwise the virus will hang around.*", as did P04: "*Oh, that is complicated. I did the same thing a while ago, but then I need to reconfigure all port forwarding again. But OK, if that helps, then we will do it again.*"

Only a few participants expressed doubts about the effort,

in all cases because they were not clear what problem Mirai posed, as with P01: "*Eh, let's take a step back. I have no idea whatsoever about how that Mirai virus actually works. I mean, I do not experience any issues, right? So what is the problem?*" After an explanation about how Mirai-infected devices are used for criminal activity against other users and organizations on the internet, P01 concluded: "*Ah, right. That is understandable, I am happy to cooperate.*". Renaud & Goucher [63] note that the 'gulf of evaluation' differentiates between the sense of being able to enact a security behaviour, and the 'response efficacy' of whether the behaviour is appropriate.

No participants dropped out before completing the steps. The only case where a participant did not want to conduct a specific step was P14, who felt none of their devices were plausible suspects, and as such did not want to implement a reset and password change on any of those devices. They did, however, proceed with subsequent steps involving the router.

Regarding the evidence for users' seemingly high motivation , one potential source of bias here (as discussed in Section 3.7) could be an observer effect (a.k.a. the 'Hawthorne effect'), where the fact that the participants know their actions are being 'observed' makes them more motivated than they might have been without the presence of the researcher.

4.5 The end: remediation, and reinfections

Table 1 presents an overview of participant-level actions and outcomes. Again, the coding used in the columns for the remediation steps relate to the boxes in Figure 2. After the intervention, 14 of the 17 participants were observed to be remediated, as measured by the absence of their IP address in the daily data feed of Mirai infections received by the ISP in the four days after the call. This may count as good news. The cumbersome non-deterministic remediation process seems at least probabilistically related to the desired outcome. Three participants remained infected. It is true that they did not fully execute the recommended steps, but the same holds for other participants who were regarded as having managed to remediate. Only four participants could be said to have fully executed the recommended actions (P01, P10, P11, P15). We include P15, because this person took the suspect device permanently offline, so in that sense 'secured' it from further harm. We monitored the presence of the IP address in the daily data feed for two more weeks after the remediation period. In 5 cases, we observed a re-infection with Mirai; there was a gap of three consecutive days where the user's IP address was not reported in the daily data feed, and then it reappeared. Two of these reinfections were non-remediated users, three were users who did manage to remediate at first.

For the two non-remediated cases, the infection disappeared by an unknown cause five or more days after the call. This is consistent with the relatively high 'natural' cleanup rate seen elsewhere [14]. One explanation is that the Mirai malware is not persistent on the device, at least not at the time of the study. This means that a power cycle may have removed the infection, although the device is still in a vulnerable state. It might be discovered and reinfected soon thereafter, because of the aggressive scanning conducted by Mirai bots.

The three cases where we observed an initial remediation, and a later reinfection, can have various explanations, and as such are indicative of avenues for future work. One explanation is that the detection of infections via the daily data feed is not perfect, potentially including false negatives. Another explanation is that these users did manage to get rid of the infections by power cycling the devices, but did not remediate the underlying vulnerability (i.e., set a secure password). This is consistent with our observations, because all three users did not fully execute the recommended actions. As noted from the observations, users may have otherwise had multiple infected devices and only focused on one, or focused their attention on the wrong device.

In the end, the gap we observed between advice and user actions cannot be blamed wholly on either the user or the advice-giver the ISP. It points to the responsibilities of a third actor: the manufacturer. Even when users went online and tried to find manufacturer information about solving security problems, there were complications. This was certainly the case for P16, who was not able to even identify the manufacturer: *"Well, there is no brand name on the device, haha, only IP-camera is printed on the side of it."*

5 Discussion

Returning to our overarching research question, we provide real-world evidence of the gap between advice and outcomes in IoT [7], but also the impact this gap can have on smart home users. There are two sides to this story – the quality of advice, and the characteristics of the response to that advice.

Successful behaviour for our participants was often unconfirmed and unconfirmable, and neither the users nor the advice-giver can resolve this at present, given the constraints inherent in the situation (in home infection, limited device visibility, etc). This unbridgable gap points to the responsibilities of other actors, notably the manufacturer [32]. We could argue users lack capability, but it is not a lack of user capability, but a design flaw, pointing to the relationship between behaviour support and interface design to provide situational feedback (as highlighted elsewhere for user access control guidance [76]). The lack of 'normal' computing interfaces on IoT devices creates an environment fraught with confusion and uncertainty for applying standard security advice.

What we have for network-connected smart home devices is also a multi-party intervention. Participants had to wait for their efforts to be confirmed as worth it (that remediation will be confirmed at some point afterwards via network scans, and a *lack* of capacity for the ISP to follow up). Participants demonstrated despair over not knowing what to do and whether their effort was successful. Remediation is then

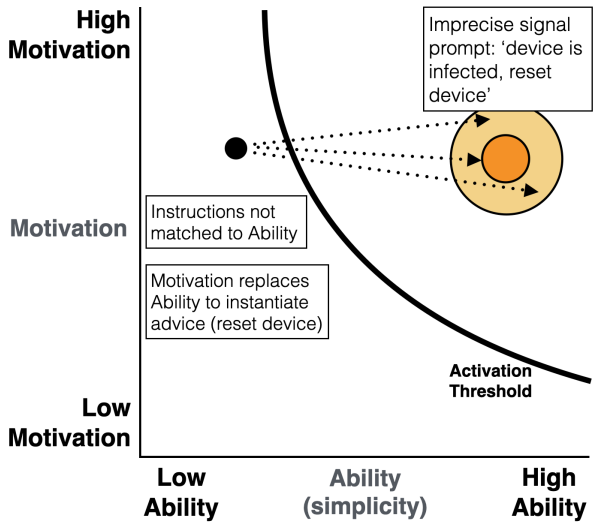


Figure 3: Action Diffraction for resetting a smart home device. Users may vary in Motivation, and rely on their Motivation to enumerate over possible solutions (standing in for a lack of knowing the precise Ability they need to apply). The target behaviour may be deterministic (the small circle, top right), but plausible variations surround it, informed in part by Instructions. It can be unclear if the applied Ability has achieved the intentions of the Prompt, even if it has been successful.

non-deterministic (very likely to work, but not definitely going to work). The lack of feedback stands in contrast to, say, removing Windows malware, where a removal tool—such as an anti-virus client—will typically report on what it found and whether it was effective in removing it. This limits the potential of checklists, for instance, if instructions cannot be made specific enough to a particular user's set of network-connected devices (and are as such, 'sub-optimally targeted').

Participants applied one of the heuristics identified in our results, to navigate the gap in specificity, and attempt to identify the target of an advocated behaviour. Applying advice then leans on motivation, in that most participants were willing to try quite convoluted steps (going to another room to unplug the router, coming back to the phone, then back to the router, etc.). Where Redmiles et al. [59] isolate 'bad advice', we step back from this to identify 'ecosystem factors which limit the capacity to construct good advice'. We regard this then as also exploring the limitations of *emergent* interventions for smart home security.

What is remarkable and worthy of further exploration is that our participants demonstrated somewhat correct reasoning in identifying suspect devices, consistent with actual properties of these devices. Mostly the heuristic is to eliminate suspect devices. This further highlights the important of local context to instantiating security advice for the smart home [76], but also making advice specific enough to be actionable [62].

5.1 Informing effective interventions

Where participants felt a need to enumerate over familiar behaviours, many would push back if they did not know how to enact the advocated behaviour. This points to self-efficacy, important for prompting action within various behaviour change approaches [24]. To put our findings in the context of enacting (what appeared to our participants as) a new behaviour, acting on notification of a malware infection is an *opportune moment* or *prompt* to enact a new behaviour, so we refer to the Fogg Behaviour Model (or B=MAP / B=MAT model [27]). In this model, Behaviour = Motivation + Ability + Prompt. The model has been used extensively across areas such as persuasive design and personal development, but also to understand social interventions for security [18], and opportunities for security interventions in a retail environment [54].

A Prompt can be a Facilitator, Spark, or Signal – here it is a Signal, that a device in the home is infected and that actions must be taken to resolve the issue, as a call to Motivation and upon an Ability to act. The ISP carries the Signal to the user (highlighting that ISPs are the *best-placed* party to intervene, but that this does not mean they are the *most appropriate*) – this relies on sufficient Motivation and Ability already being present. We found that participants were *over-investing* Motivation to make up for an insufficient definition of the target behaviour or outcome (a lack of capability to identify or confirm the appropriate Ability). Among our participants, there was uncertainty as to what was right to do, to the extent that a user may enact a behaviour which removes malware, but continue with further actions for lack of indication that they had already succeeded. This even includes where some of our participants chose to permanently disconnect or dispose of a suspect device (representing an *unintended harm* of unclear advice [16]). 'Actionable choices', with clear outcomes, are regarded as feasible in areas such as smart home privacy [66], and in supporting a user-defined 'recovery state' [35].

We show the gulf where these harms manifest as what we refer to here as 'Action Diffraction' (Figure 3). Where Renaud & Goucher refer to the 'Gulf of Execution' [63] (including knowing what needs to be done, but not how to do it), here we find a gulf created by restrictions in the vehicle of the intervention itself which makes the target behaviour indistinct. This applies to both knowing what the target behaviour is, and knowing whether it has been reached. Where the Activation Threshold is the point of realising a target Behaviour, and a user being activated to try to get over the Threshold, our results show efforts being 'diffracted', splitting off in many directions as participants find themselves exploring non-deterministic and potentially inapplicable behaviours (this includes where they have Ability to do something, but are not willing to try everything unless they can be Motivated to do so).

Renaud & Goucher [63] frame a 'Gulf of Evaluation' in formulating an intention to adopt a secure behaviour, and Redmiles et al. [61] identify dimensions of advice quality.

We note in reference to the latter that the specificity – and actionability – of advice, including the capacity to evaluate the efficacy of the behaviour [63], are also impacted by the specificity of the target behaviour and its confirmation. Our findings showed also that, as with other forms of security advice [62], multiple sources of instructions can potentially confuse users further.

One contributory factor to this problem is best articulated through the Behaviour Wizard of Fogg & Hreha [28]. The best-practice advice seen by smart home users is an 'unfamiliar' task (requiring a link to existing practices), but framed more like a 'familiar' task (one that does not need explanation), and so we saw participants replacing an unfamiliar action with familiar behaviour(s). This is a complex world of Things, where enacting the wrong behaviour can result in 'proxy changes' [53], regardless of whether the intended outcome is reached. A user may turn on and off many devices, or the wrong one and not the right one, or achieve the goal but lose tailored configuration settings in the effort , all while not knowing in the moment whether they have succeeded.

5.2 Implications for evolving IoT threats

If users only apply some of the advice they are given, or devices have inherent security weaknesses, they may *continue* to be vulnerable and require *regular* intervention. Users can follow advice but still suffer the same consequences again, if IoT infrastructure does not help them to stay recovered, or malware evolves. There are parallels to the Transtheoretical Model [57], where understanding specific stages of behaviour can identify security improvements [55]. Inherent weaknesses in the design of many smart home devices put a user back into an 'unhealthy' situation (e.g., a device repeatedly falling back into an insecure state), requiring repeated cycles of *contemplating* and *acting* on advice, to *maintain* secure devices.

New malware variants are moving away from short-lived infections, and becoming persistent and resistant to current interventions [12]. More efforts of the type we have observed for Mirai infections would be required where, for instance, thousands of QNAP network access storage devices have been targeted by persistent malware [75], and the direction of travel shows that advice from manufacturers is requiring users to follow 20 or more steps completely and successfully to resolve these issues [58]. Moreover, some of these variants are also starting to include countermeasures to make detection difficult. For instance, malware leveraging blockchain DNS or TOR makes it even harder for the interveners to assess the efficacy of the user's actions [10, 69, 73]. This is all within the context of increasing use of smart home devices, which itself already increases the complexity of remediation when there are problems (as we saw evidence of here).

5.3 Recommendations

Here we describe recommendations emerging from our Results and consideration of behaviour change approaches, associating recommendations to specific stakeholders.

- **Confirmation of settings changes.** Visibility of changes to system status is a crucial design principle [52]. Here this applies to both Apps and Interfaces, as created and maintained by the *manufacturer*. This was seen among our participants as already happening for some devices and interfaces, but should be enshrined as a consistent design choice, to reduce the 'diffraction' of remediation efforts. This would then serve as a visible 'security outcome' [29], to *then* be able to consider whether the visible outcome was the correct step to follow. This may be necessary for future security issues if resetting / unplugging a device actually runs the risk of reinstating default credentials, for instance. This would complement efforts to standardise smart home device functionality (as in e.g., the UK [70] and US [26]) which aim to have *manufacturers* reduce the scope for misconfiguration as a vector for device compromise (as with e.g., easily-guessed 'default' settings).

- **Settings logs.** A log of settings changes can help both *users* and *ISPs* (or indeed anyone 'helping' users) to see and refer to a clear record of changes. This could also include notifying users of security settings which need to be changed at setup but have not been, or which have been changed but not by a registered user. Ideally, there would be some signalling to users when a security issue is suspected, where there is a general lack of event logging related to security [26].

- **Assisted remediation.** Our study showed that not all participants were able to follow the advice, or needed confirmation that they had followed it. For lack of being able to move incrementally toward a clearly focused outcome (Figure 3), having a helper on-call or on-site would increase chances of a successful outcome, if the previous steps cannot be achieved. This would be a low-bar in terms of ensuring that there is an intervention for all levels of Motivation and Ability – if *users* are as keen to follow advice as our participants, they cannot be blamed if they are trapped in a cycle of trying advice without confirmation of actions or visible evidence of success. This relates to having actionable choices to begin with. It also aligns with the incentives of ISPs, which could commercially offer such services, though this brings the risk of users distrusting notifications as a ploy to sell a service – ISPs might only offer the service if the user asks for it.

6 Related Work

Chalhoub & Flechais [15] studied real-world users of smart home devices, where limitations in device features and transparency were seen to frustrate privacy-related decisions. The authors characterised *compensatory behaviours* in response to concerns (such as disconnecting a device). We saw participants defaulting to 'familiar' behaviours as a strategy to approach the uncertain process of situating generic advice. Geeng & Roesner [31] studied multi-user smart homes, noting that when devices fail to function properly, alternative paths to using a device are needed. We saw a parallel, where participants sought a viable solution to critical security issues, but were at times reluctant to dismantle their smart home device configurations to achieve it.

In terms of supporting behaviour change, Forget et al. [29] studied the security attitudes, behaviours, and understanding of active computer users from device activity and interviews. The authors characterised ineffectively proactive users, who exerted too much effort for security or regularly performed familiar behaviours even if they did not match the security concern. Where the authors saw information-seeking behaviours, our participants felt challenged in determining what to seek information about (lacking both clarity as to what was the target device, and available diagnostic information). Crucially, Forget et al. highlighted the importance of tangible outcomes to user actions, where here there was a lack of clear outcomes; the authors identified 'problematic knowledge gaps', where for consumer IoT environments these gaps are constraints in advice and user support.

Reeder et al. [62] identify a range of criteria for good home security advice, including that it must be actionable. We identify a gap that requires the recipient of smart home security advice to be able to complete advice and relate advice received from others to their personal context. The authors also discuss the potential need to enumerate over possible versions of generic advice to reach specific advice, considering "offering the generic advice followed by specific instructions on how to implement it" – similarly, our participants applied strategies to do this themselves.

Redmiles et al. [61] identify 'perceived efficacy' of advice as important, where here there is an element of efficacy in being able to localise advice received from others. The advice the authors reviewed was regarded as mostly 'actionable', where here we explore the implications of advice which, at least for our participants, was not immediately actionable. Redmiles et al. regard network security as amongst the least actionable and most general security advice (e.g., "Secure your router"), raising questions of whether non-actionable advice should be given to users in the first place, and we provide real-world evidence informing this discussion.

Çetin et al. [13] studied a 'walled garden' approach of limiting users' capacity to access the Internet while a device is infected. Here we learned about the remediation journey while users were acting on suggested remediation actions locally themselves, rather than checking the effectiveness of the notification method alone.

7 Conclusion

Here we studied user efforts to apply advice provided to them by their ISP. We found that the advice was not specific enough to ensure that it was applicable to participants' own smart home context. Critically, constraints to the specificity of advice limited how it was produced, communicated, and put into practice in a real-world setting. Only 4 of 17 participants completed all applicable advice steps successfully. Action typically went wrong at the second step (changing the password of the suspected device), or at the fourth step (resetting the router to its factory settings). 16 participants were able to pinpoint a plausible infected device, using one of three strategies we identified (including by process of elimination).

Our work informs the understanding of interventions for real-world IoT settings. The construction, communication, and enactment of technical advice to home users is both complex and collaborative. It involves end-users, their ISPs, device manufacturers, and technical experts to support successful outcomes. Putting our findings into perspective with the continuing need for technical support for home computers and mobile devices, the need to fix security issues of smart home devices can be expected to persist. Given the complexity and role of local context, this can be expected to require analysis of the smart home in situ, including return visits to users of reinfected devices. Future work will explore the capacity of intervention approaches which include multiple relevant stakeholders. For instance, a list of known vulnerable device models could aid both ISPs in informing end-users, and end-users themselves in identifying problematic devices which they use or are considering for purchase.

Acknowledgments

We thank the partner ISP company for permitting access to network data and knowledgeable staff, and facilitating engagement with their customers. This publication is part of the MINIONS project (number 628.001.033) of the "Joint U.S.-Netherlands Cyber Security Research Programme" which is (partly) financed by the Dutch Research Council (NWO). We also wish to thank our paper reviewers for their comments.

References

[1] Fadele Ayotunde Alaba, Mazliza Othman, Ibrahim Abaker Targio Hashem, and Faiz Alotaibi. Internet of Things security: A survey. *Journal of Network and Computer Applications*, 88:10–28, 2017.

[2] D Allen. Transaction costs. *Encyclopedia of Law and Economics*, 1999.

[3] Omar Alrawi, Chaz Lever, Manos Antonakakis, and Fabian Monrose. SoK: Security evaluation of home-based IoT deployments. In *2019 IEEE symposium on Security and Privacy (S&P)*, pages 1362–1380. IEEE, 2019.

[4] Rosemarie Anderson. Thematic content analysis (TCA). *Descriptive presentation of qualitative data*, pages 1–4, 2007.

[5] Eirini Anthi, Shazaib Ahmad, Omer Rana, George Theodorakopoulos, and Pete Burnap. EclipseIoT: A secure and adaptive hub for the Internet of Things. *Computers & Security*, 78:477–490, 2018.

[6] Manos Antonakakis, Tim April, Michael Bailey, Matt Bernhard, Elie Bursztein, Jaime Cochran, Zakir Durumeric, J Alex Halderman, Luca Invernizzi, Michalis Kallitsis, et al. Understanding the Mirai botnet. In *26th USENIX security symposium (USENIX Security '17)*, pages 1093–1110, 2017.

[7] Christopher Bellman and Paul C van Oorschot. Best practices for IoT security: What does that even mean? *arXiv preprint arXiv:2004.12179*, 2020.

[8] JM Blythe and SD Johnson. The consumer security index for IoT: A protocol for developing an index to improve consumer decision making and to incentivize greater security provision in IoT devices. *Living in the Internet of Things: Cybersecurity of the IoT*, 2018.

[9] John M Blythe, Shane D Johnson, and Matthew Manning. What is security worth to consumers? investigating willingness to pay for secure Internet of Things devices. *Crime Science*, 9(1):1–9, 2020.

[10] Leon Böck, Nikolaos Alexopoulos, Emine Saracoglu, Max Mühlhäuser, and Emmanouil Vasilomanolakis. Assessing the threat of blockchain-based botnets. In *2019 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–11. IEEE, 2019.

[11] Irina Brass, Leonie Tanczer, Madeline Carr, Miles Elsden, and Jason Blackstock. Standardising a moving target: The development and evolution of IoT security standards. *IET*, 2018.

[12] Calvin Brierley, Jamie Pont, Budi Arief, David J Barnes, and Julio C Hernandez-Castro. Persistence in Linux-based IoT malware. *NordSec 2020*, 2020.

[13] Orçun Çetin, Lisette Altena, Carlos Gañán, and Michel Van Eeten. Let me out! Evaluating the effectiveness of quarantining compromised users in walled gardens. *Fourteenth Symposium on Usable Privacy and Security (SOUPS)*, 2018.

[14] Orçun Çetin, Carlos Gañán, Lisette Altena, Takahiro Kasama, Daisuke Inoue, Kazuki Tamiya, Ying Tie, Katsunari Yoshioka, and Michel Van Eeten. Cleaning up the internet of evil things: Real-world evidence on ISP and consumer efforts to remove Mirai. In *Network and Distributed System Security Symposium (NDSS)*, 2019.

[15] George Chalhoub and Ivan Flechais. "Alexa, are you spying on me?": Exploring the effect of user experience on the security and privacy of smart speaker users. In *International Conference on Human-Computer Interaction*, pages 305–325. Springer, 2020.

[16] Yi Ting Chua, Simon Parkin, Matthew Edwards, Daniela Oliveira, Stefan Schiffner, Gareth Tyson, and Alice Hutchings. Identifying unintended harms of cybersecurity countermeasures. In *2019 APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–15. IEEE, 2019.

[17] Andrei Costin, Jonas Zaddach, Aurélien Francillon, and Davide Balzarotti. A large-scale analysis of the security of embedded firmwares. In *23rd USENIX Security Symposium (USENIX Security '14)*, pages 95–110, 2014.

[18] Sauvik Das, Laura A Dabbish, and Jason I Hong. A typology of perceived triggers for end-user security and privacy behaviors. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*, 2019.

[19] Michele De Donno, Nicola Dragoni, Alberto Giaretta, and Angelo Spognardi. DDoS-capable IoT malwares: Comparative analysis and Mirai investigation. *Security and Communication Networks*, 2018, 2018.

[20] David Dittrich, Erin Kenneally, et al. The Menlo report: Ethical principles guiding information and communication technology research. http://www.caida.org/publications/papers/2012/menlo_report_actual_formatted, 2012. Accessed:2021-05-25.

[21] Zakir Durumeric, Frank Li, James Kasten, Johanna Amann, Jethro Beekman, Mathias Payer, Nicolas Weaver, David Adrian, Vern Paxson, Michael Bailey, et al. The matter of Heartbleed. In *Proceedings of the 2014 conference on internet measurement conference*, pages 475–488, 2014.

[22] Pardis Emami-Naeini, Yuvraj Agarwal, Lorrie Faith Cranor, and Hanan Hibshi. Ask the experts: What should be on an IoT privacy and security label? In *2020 IEEE Symposium on Security and Privacy (S & P)*, pages 447–464. IEEE, 2020.

[23] European Union & Agency for Network and Information Security. Baseline security recommendations for IoT in the context of critical information infrastructures - Publications Office of the EU. https://op.europa.eu/en/publication-detail/-/publication/c37f8196-d96f-11e7-a506-01aa75ed71a1/language-en, 2017. Accessed:2021-05-25.

[24] European Union Agency for Cybersecurity (ENISA). Cybersecurity culture guidelines: Behavioural aspects of cybersecurity. https://www.enisa.europa.eu/publications/cybersecurity-culture-guidelines-behavioural-aspects-of-cybersecurity, 2018. Accessed:2021-05-25.

[25] European Union and Council of Europe. Document Library | Europass. https://europa.eu/europass/en/document-library, 2004. Accessed: 2021-05-25.

[26] Michael Fagan, Mary Yang, Allen Tan, Lora Randolph, and Karen Scarfone. Security review of consumer home Internet of Things (IoT) products. https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8267-draft.pdf, 2019.

[27] Brian J Fogg. A behavior model for persuasive design. In *Proceedings of the 4th international Conference on Persuasive Technology*, pages 1–7, 2009.

[28] Brian J Fogg and Jason Hreha. Behavior wizard: A method for matching target behaviors with solutions. In *International Conference on Persuasive Technology*, pages 117–131. Springer, 2010.

[29] Alain Forget, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, Marian Harbach, and Rahul Telang. Do or do not, there is no try: user engagement may not improve security outcomes. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 97–111, 2016.

[30] Steven Furnell. Making security usable: Are things improving? *Computers & Security*, 26(6):434–443, 2007.

[31] Christine Geeng and Franziska Roesner. Who's in control? interactions in multi-user smart homes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[32] Julie M Haney, Yasemin Acar, and Susanne M Furman. "it's the company, the government, you and I": User perceptions of responsibility for smart home privacy and security. In *30th USENIX Security Symposium (USENIX Security '21)*, Vancouver, B.C., August 2021. USENIX Association.

[33] Sam Havron, Diana Freed, Rahul Chatterjee, Damon McCoy, Nicola Dell, and Thomas Ristenpart. Clinical computer security for victims of intimate partner violence. In *28th USENIX Security Symposium (USENIX Security '19)*, pages 105–122, 2019.

[34] Weijia He, Maximilian Golla, Roshni Padhi, Jordan Ofek, Markus Dürmuth, Earlence Fernandes, and Blase Ur. Rethinking access control and authentication for the home internet of things (IoT). In *27th USENIX Security Symposium (USENIX Security '18)*, pages 255–272, 2018.

[35] Weijia He, Jesse Martinez, Roshni Padhi, Lefan Zhang, and Blase Ur. When smart devices are stupid: negative experiences using home smart devices. In *2019 IEEE Security and Privacy Workshops (SPW)*, pages 150–155. IEEE, 2019.

[36] Bonnie E John and Steven J Marks. Tracking the effectiveness of usability evaluation methods. *Behaviour & Information Technology*, 16(4-5):188–202, 1997.

[37] Shane D Johnson, John M Blythe, Matthew Manning, and Gabriel TW Wong. The impact of IoT security labelling on consumer product choice and willingness to pay. *PloS one*, 15(1):e0227800, 2020.

[38] G. Kambourakis, C. Kolias, and A. Stavrou. The Mirai botnet and the IoT zombie armies. In *MILCOM 2017 - 2017 IEEE Military Communications Conference (MILCOM)*, pages 267–272, 2017.

[39] Kaspersky. New Mirai botnet is targeting enterprise IoT | Kaspersky official blog. https://www.kaspersky.com/blog/mirai-enterprise/26032/, 2019. Accessed: 2021-05-25.

[40] Constantinos Kolias, Georgios Kambourakis, Angelos Stavrou, and Jeffrey Voas. DDoS in the IoT: Mirai and other botnets. *Computer*, 2017.

[41] David Kotz and Travis Peters. Challenges to ensuring human safety throughout the life-cycle of smart environments. In *Proceedings of the 1st ACM Workshop on the Internet of Safe Things*, pages 1–7, 2017.

[42] Clayton Lewis. *Using the "thinking-aloud" method in cognitive interface design*. IBM TJ Watson Research Center Yorktown Heights, NY, 1982.

[43] Frank Li, Grant Ho, Eric Kuan, Yuan Niu, Lucas Ballard, Kurt Thomas, Elie Bursztein, and Vern Paxson. Remedying web hijacking: Notification effectiveness and webmaster comprehension. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1009–1019, 2016.

[44] Jason Livingood, Nirmal Mody, and Mike O'Reirdan. Recommendations for the Remediation of Bots in ISP Networks. https://tools.ietf.org/html/rfc6561, 2012. Accessed:2021-05-25.

[45] J. Margolis, T. T. Oh, S. Jadhav, Y. H. Kim, and J. N. Kim. An in-depth analysis of the Mirai botnet. In *2017 International Conference on Software Security and Assurance (ICSSA)*, pages 6–12, 2017.

[46] David Moore. Network telescopes: Observing small or distant security events. In *11th USENIX Security Symposium (USENIX Security '02)*, San Francisco, CA, August 2002. USENIX Association.

[47] Philipp Morgner, Christoph Mai, Nicole Koschate-Fischer, Felix Freiling, and Zinaida Benenson. Security update labels: Establishing economic incentives for security patching of IoT consumer products. In *2020 IEEE Symposium on Security and Privacy (S&P)*, pages 429–446. IEEE, 2020.

[48] Mozilla. *Privacy not included. https://foundation.mozilla.org/en/privacynotincluded/, 2021. Accessed: 2021-05-25.

[49] Netscout. Dawn of the terrorbit era. https://www.netscout.com/sites/default/files/2019-02/SECR_001_EN-1901%20-%20NETSCOUT%20Threat%20Intelligence%20Report%202H%202018.pdf, 2018. Accessed: 2021-05-25.

[50] Kenneth D Nguyen, Heather Rosoff, and Richard S John. Valuing information security from a phishing attack. *Journal of Cybersecurity*, 3(3):159–171, 2017.

[51] Larissa Nicholls, Yolande Strengers, and Jathan Sadowski. Social impacts and control in the smart home. *Nature Energy*, 5(3):180–182, 2020.

[52] Jakob Nielsen. Enhancing the explanatory power of usability heuristics. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 152–158, 1994.

[53] Magda Osman, Scott McLachlan, Norman Fenton, Martin Neil, Ragnar Löfstedt, and Björn Meder. Learning from behavioural changes that fail. *Trends in Cognitive Sciences*, 2020.

[54] Simon Parkin, Elissa M Redmiles, Lynne Coventry, and M Angela Sasse. Security when it is welcome: Exploring device purchase as an opportune moment for security behavior change. In *Proceedings of the Workshop on Usable Security and Privacy (USEC'19)*. Internet Society, 2019.

[55] Shari Lawrence Pfleeger, M Angela Sasse, and Adrian Furnham. From weakest link to security hero: Transforming staff security behavior. *Journal of Homeland Security and Emergency Management*, 11(4):489–510, 2014.

[56] Erika Shehan Poole, Marshini Chetty, Tom Morgan, Rebecca E Grinter, and W Keith Edwards. Computer help at home: methods and motivations for informal technical support. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 739–748, 2009.

[57] James O Prochaska, Colleen A Redding, Kerry E Evers, et al. The transtheoretical model and stages of change. *Health behavior: Theory, research, and practice*, 97, 2015.

[58] QNAP. Security Advisory for Malware QSnatch - Security Advisory | QNAP. https://www.qnap.com/en/security-advisory/nas-201911-01, 2021. Accessed: 2021-05-25.

[59] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. Where is the digital divide? a survey of security, privacy, and socioeconomics. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 931–936, 2017.

[60] Elissa M Redmiles, Amelia R Malone, and Michelle L Mazurek. I think they're trying to tell me something: Advice sources and selection for digital security. In *2016 IEEE Symposium on Security and Privacy (S&P)*, pages 272–288. IEEE, 2016.

[61] Elissa M. Redmiles, Noel Warford, Amritha Jayanti, Aravind Koneru, Sean Kross, Miraida Morales, Rock Stevens, and Michelle L. Mazurek. A comprehensive quality evaluation of security and privacy advice on the web. In *29th USENIX Security Symposium (USENIX Security '20)*, pages 89–108. USENIX Association, August 2020.

[62] Robert W Reeder, Iulia Ion, and Sunny Consolvo. 152 simple steps to stay safe online: Security advice for non-tech-savvy users. *IEEE Security & Privacy*, 15(5):55–64, 2017.

[63] Karen Renaud and Wendy Goucher. The curious incidence of security breaches by knowledgeable employees and the pivotal role a of security culture. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*, pages 361–372. Springer, 2014.

[64] Brent Rowe and Dallas Wood. Are home internet users willing to pay ISPs for improvements in cyber security? In *Economics of information security and privacy III*, pages 193–212. Springer, 2013.

[65] Bruce Schneier. *Click here to kill everybody: Security and survival in a hyper-connected world*. WW Norton & Company, 2018.

[66] William Seymour, Martin J Kraemer, Reuben Binns, and Max Van Kleek. Informing the design of privacy-empowering tools for the connected home. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[67] Shadowserver. Drone/Botnet-Drone Report | Shadowserver. https://www.shadowserver.org/what-we-do/network-reporting/drone-botnet-drone-report/, 2021. Accessed: 2021-05-25.

[68] H. Sinanović and S. Mrdovic. Analysis of Mirai malicious software. In *2017 25th International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, pages 1–5, 2017.

[69] Alex Turing, Hui Wang, and Liu Yang. New threat: Matryosh botnet is spreading. https://blog.netlab.360.com/matryosh-botnet-is-spreading-en/, 2021. Accessed: 2021-05-25.

[70] UK Department for Digital, Culture, Media & Sport (DCMS). Code of practice for consumer IoT security. https://www.gov.uk/government/publications/code-of-practice-for-consumer-iot-security, 2018.

[71] Maaike Van Den Haak, Menno De Jong, and Peter Jan Schellens. Retrospective vs. concurrent think-aloud protocols: testing the usability of an online library catalogue. *Behaviour & information technology*, 22(5):339–351, 2003.

[72] Marie Vasek and Tyler Moore. Do malware reports expedite cleanup? an experimental study. In *5th Workshop on Cyber Security Experimentation and Test (CSET '12)*, 2012.

[73] Hui Wang. Fbot, a Satori related botnet using block-chain DNS system. https://blog.netlab.360.com/threat-alert-a-new-worm-fbot-cleaning-adbminer-is-using-a-blockchain-based-dns-en/, 2018.

[74] Rick Wash, Emilee Rader, Kami Vaniea, and Michelle Rizor. Out of the loop: How automated software updates cause unintended security consequences. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 89–104, 2014.

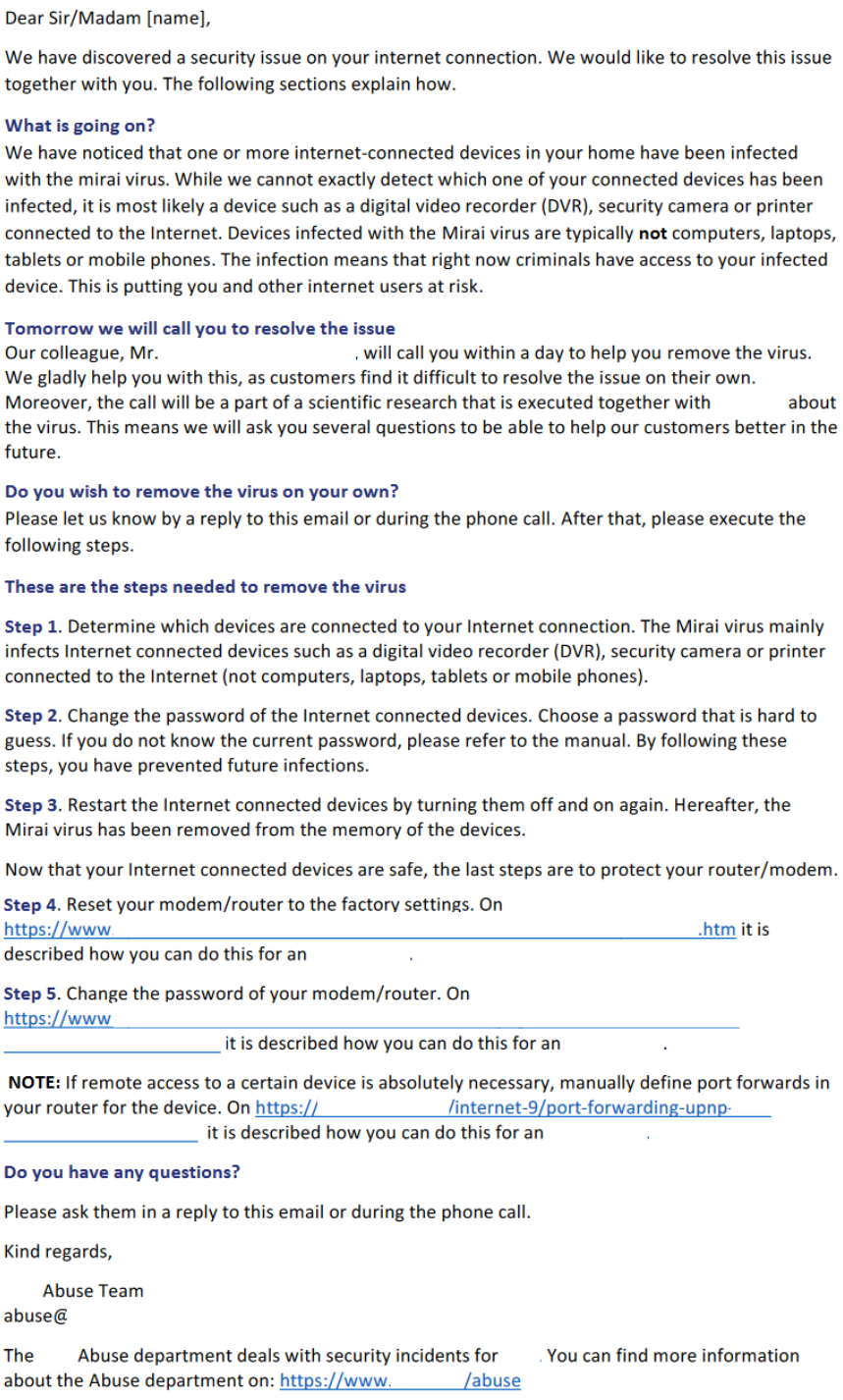
[75] ZDNet. Thousands of QNAP NAS devices have been infected with the QSnatch malware | ZDNet. https://www.zdnet.com/article/thousands-of-qnap-nas-devices-have-been-infected-with-the-qsnatch-malware/, 2019. Accessed: 2021-05-25.

[76] Eric Zeng, Shrirang Mare, and Franziska Roesner. End user security and privacy concerns with smart homes. In *thirteenth symposium on usable privacy and security (SOUPS 2017)*, pages 65–80, 2017.

[77] Eric Zeng and Franziska Roesner. Understanding and improving security and privacy in multi-user smart homes: a design exploration and in-home user study. In *28th USENIX Security Symposium (USENIX Security '19)*, pages 159–176, 2019.

[78] Verena Zimmermann, Paul Gerber, Karola Marky, Leon Böck, and Florian Kirchbuchner. Assessing users' privacy and security concerns of smart home technologies. *i-com*, 18(3):197–216, 2019.

A Appendix A – Notification Message and Instructions



Dear Sir/Madam [name],

We have discovered a security issue on your internet connection. We would like to resolve this issue together with you. The following sections explain how.

What is going on?

We have noticed that one or more internet-connected devices in your home have been infected with the mirai virus. While we cannot exactly detect which one of your connected devices has been infected, it is most likely a device such as a digital video recorder (DVR), security camera or printer connected to the Internet. Devices infected with the Mirai virus are typically **not** computers, laptops, tablets or mobile phones. The infection means that right now criminals have access to your infected device. This is putting you and other internet users at risk.

Tomorrow we will call you to resolve the issue

Our colleague, Mr. , will call you within a day to help you remove the virus. We gladly help you with this, as customers find it difficult to resolve the issue on their own. Moreover, the call will be a part of a scientific research that is executed together with about the virus. This means we will ask you several questions to be able to help our customers better in the future.

Do you wish to remove the virus on your own?

Please let us know by a reply to this email or during the phone call. After that, please execute the following steps.

These are the steps needed to remove the virus

Step 1. Determine which devices are connected to your Internet connection. The Mirai virus mainly infects Internet connected devices such as a digital video recorder (DVR), security camera or printer connected to the Internet (not computers, laptops, tablets or mobile phones).

Step 2. Change the password of the Internet connected devices. Choose a password that is hard to guess. If you do not know the current password, please refer to the manual. By following these steps, you have prevented future infections.

Step 3. Restart the Internet connected devices by turning them off and on again. Hereafter, the Mirai virus has been removed from the memory of the devices.

Now that your Internet connected devices are safe, the last steps are to protect your router/modem.

Step 4. Reset your modem/router to the factory settings. On https://www .htm it is described how you can do this for an .

Step 5. Change the password of your modem/router. On https://www it is described how you can do this for an .

NOTE: If remote access to a certain device is absolutely necessary, manually define port forwards in your router for the device. On https:// /internet-9/port-forwarding-upnp it is described how you can do this for an .

Do you have any questions?

Please ask them in a reply to this email or during the phone call.

Kind regards,

Abuse Team
abuse@

The Abuse department deals with security incidents for . You can find more information about the Abuse department on: https://www. /abuse

Figure 4: Notification and opt-out invitation

B Appendix B – Think-Aloud Protocol

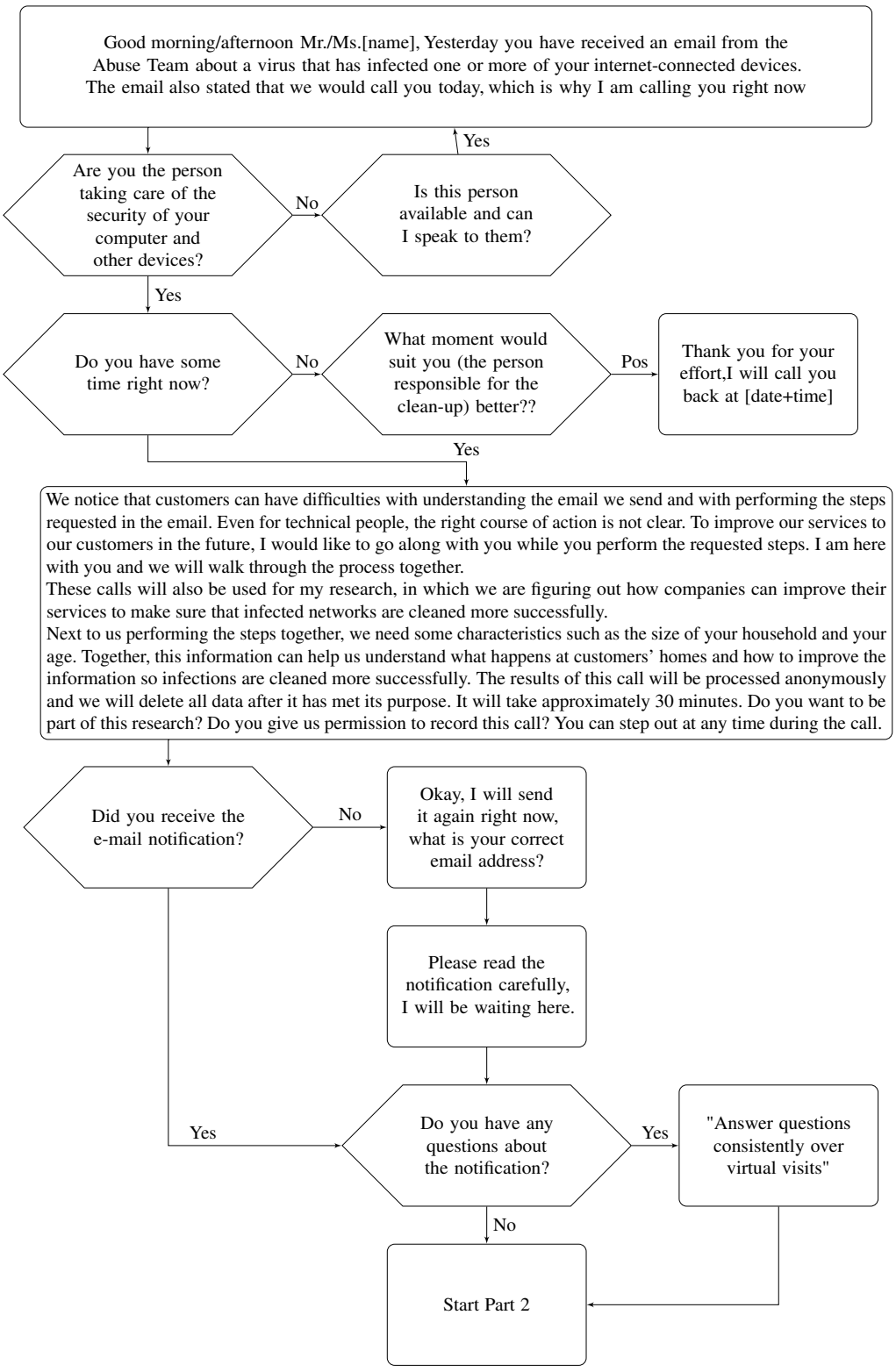


Figure 5: Think-aloud protocol - Part 1

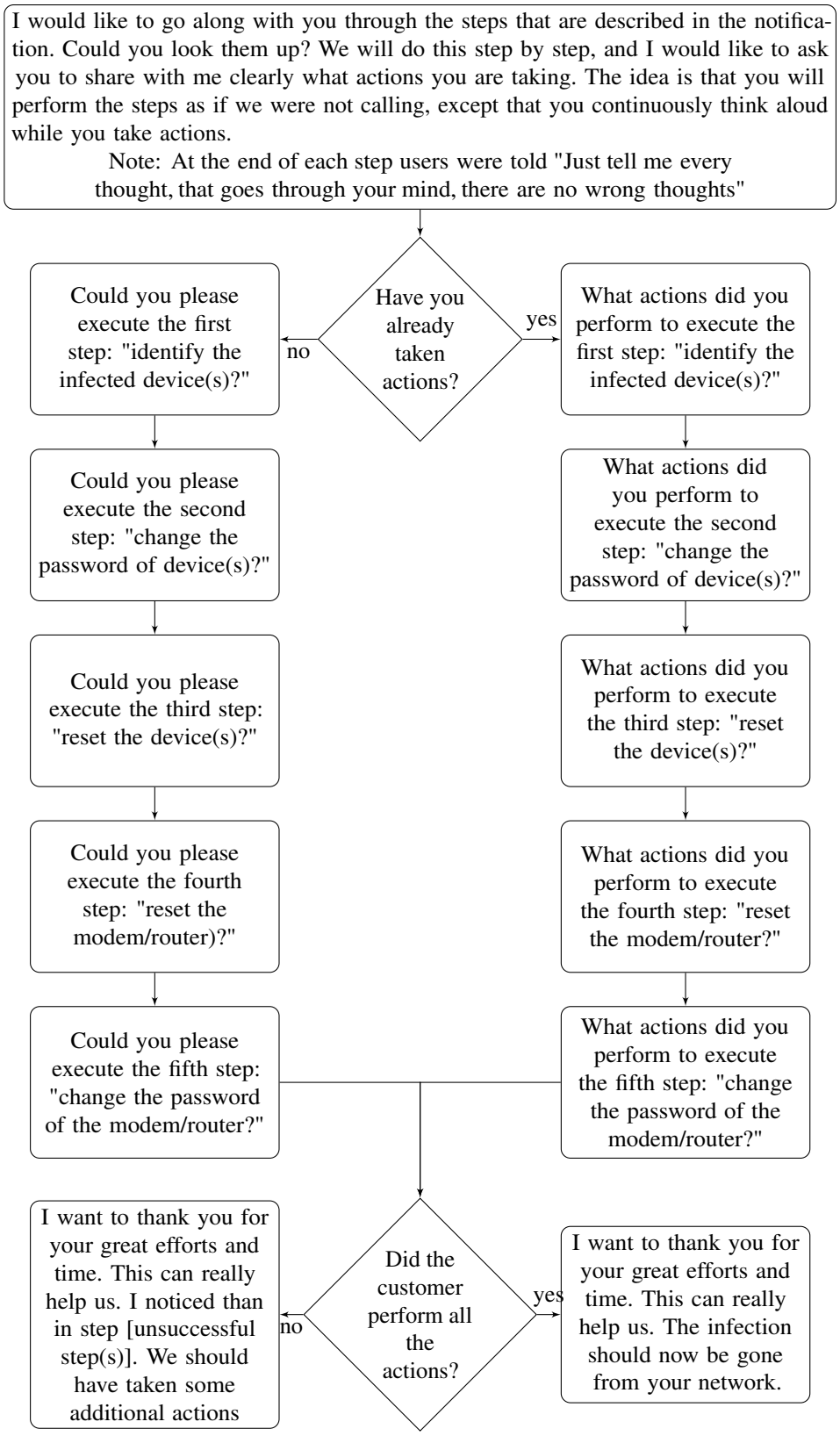


Figure 6: Think-aloud protocol - Part 2

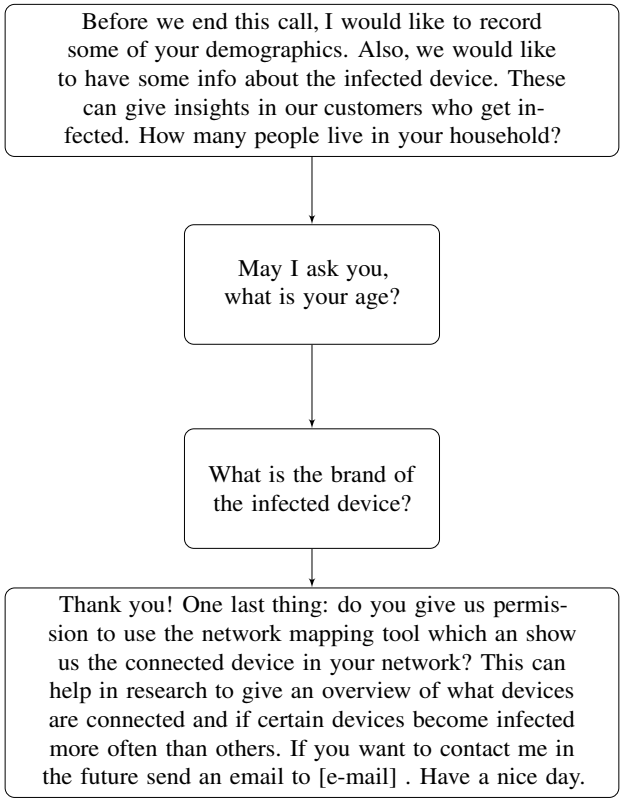


Figure 7: Think-aloud protocol - Part 3

Evaluating and Redefining Smartphone Permissions with Contextualized Justifications for Mobile Augmented Reality Apps

David Harborth
Goethe University Frankfurt am Main

Alisa Frik
ICSI, University of California Berkeley

Abstract

Augmented reality (AR), and specifically mobile augmented reality (MAR) gained much public attention after the success of Pokémon Go in 2016, and since then has found application in online games, social media, entertainment, real estate, interior design, and other services. MAR apps are highly dependent on real time context-specific information provided by the different sensors and data processing capabilities of smartphones (e.g., LiDAR, gyroscope or object recognition). This dependency raises crucial privacy issues for end users. We evaluate whether the existing access permission systems, initially developed for non-AR apps, as well as proposed new permissions, relevant for MAR apps, provide sufficient and clear information to the users. We address this research goal in two online survey-based experiments with a total of 581 participants. Based on our results, we argue that it is necessary to increase transparency about MAR apps' data practices by requesting users' permissions to access certain novel and privacy invasive resources and functionalities commonly used in MAR apps, such as speech and face recognition. We also find that adding justifications, contextualized to the data collection practices of the app, improves transparency and can mitigate privacy concerns, at least in the context of data utilized to the users' benefit. Better understanding of the app's practices and lower concerns, in turn, increase the intentions to grant permissions. We provide recommendations for better transparency in MAR apps.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

1 Introduction

The release of Pokémon Go in 2016 increased the public awareness about augmented reality (AR) [41]. AR is defined as a technology which "combines real and virtual objects in a real environment; runs interactively, and in real time; and registers (aligns) real and virtual objects with each other" [4, p.34]. The AR market in general was worth $1.8 billion in 2018, $3.5 billion in 2019 and is expected to increase in value to $18 billion by 2023 [10]. Almost a quarter of the US population, 72.8 million people, used AR at least once a month in 2019. It was projected to increase to 83.1 million people in 2020 (representing 25.3% of the US population) [43].

Currently, the two most popular types of AR are smart glasses and mobile AR (MAR) apps. AR glasses such as the Microsoft Hololens [38] are not yet mature enough products for the end consumer market due to the large weight and size, and high price. This type of AR is primarily used in the Business-to-Business (B2B) environment in which AR could successfully demonstrate its value by saving time and money in numerous processes [28]. However, recent news reveal that Apple is planning to release AR glasses in the near future, which could lead to a major breakthrough of AR smart glasses in the end consumer market [29]. In contrast, MAR apps and AR features within regular mobile apps are already widely available and used within the smartphone ecosystem. One of the most famous examples is the aforementioned MAR game Pokémon Go — one of the most successful mobile apps ever introduced, which generated $1.8 billion revenue in two years [40]. Other popular apps like Snapchat and Instagram integrate AR filters as well, which became even more popular during the COVID-19 pandemic's lockdown among users now spending more of their time in smartphones [52].

In order to create engaging interactive experience, MAR apps require large amounts of data from a variety of sensors, processed using machine learning and artificial intelligence algorithms (e.g., for object recognition, and geometry tracking). Such data-intensive processes inevitably raise concerns about user privacy and security. Based on the analysis of prior

literature, we identify five major differences between MAR apps and non-MAR apps, which amplify privacy and security risks for MAR app users [9, 21]:

1. heavy reliance on the camera input but limited feedback regarding what data is captured by the camera and used by the MAR app;
2. malicious apps can realistically alter digital objects and information presented to the user and deceive them;
3. increased data aggregation capabilities of MAR apps when combining the output of multiple sensors (e.g., location, visual/camera, accelerometer data, etc.), and the opacity of potential inferences and risks associated with such aggregation to the user;
4. privacy breaches in collaborative and shared MAR environments when two or more users work on the same digital objects using separate devices [32];
5. bystanders of MAR systems who are in the field of view and get filmed by the systems without awareness or possibility to control [11].

Therefore, users of MAR apps are exposed to more severe and novel types of privacy risks compared to the ones related to regular, non-MAR, smartphones apps. However, there is a lack of user studies investigating MAR related privacy concerns of users [12, 19]. We contribute research in that field.

While in general data flows in the apps are not transparent to the users [5, 13, 24], a few most common ways in which apps' data collection practices are revealed to the user is through the permission systems and privacy policies. However, not all apps provide privacy policies [47, 49], and even when they do, they are hard to understand for non-expert users which decreases the likelihood that they read the policies [7, 44]. Thus, despite criticism related to a partially ineffective design [17, 26, 56], permissions remain an integral and mandatory element for collecting app user's data. Therefore, we decided to start our analysis of MAR app users' privacy concerns with an investigation of their opinions about the existing permission systems in order to provide recommendations for increased transparency.

Our study addresses the following research questions:

RQ1: How does information provided in smartphone permissions affect users' understanding of what resources and data are accessed by an MAR app? And what are users' expectations about the impact of these permissions on the app's performance?

RQ2: How does information provided in permissions affect users' privacy concerns regarding MAR apps?

RQ3: How do the justifications for requesting smartphone permissions affect users' choices regarding whether to grant such permissions and whether to download an MAR app?

RQ4: How can the transparency of smartphone permissions in MAR apps be improved?

To address these research questions, we conducted two online survey-based experiments with a total of 581 participants (both studies were approved by the university's ethics board). We explored their understanding of a hypothetical MAR app's data practices based on the permissions it requests. We tested both the existing permissions (e.g., to access contacts and camera) and the proposed new ones that are currently not requested in mobile apps, but are commonly accessed by MAR apps without permission, despite having serious privacy implications (e.g., LiDAR, accelerometer and gyroscope, object recognition, etc.). We also added the justifications about how the app will use the data should the permission be granted. In this study, we limited the purposes of data use to the ones relevant to the app's functionalities and overall beneficial to the users (as opposed to malicious or non beneficial data use). We tested whether in addition to the currently used permission labels, the inclusion of such justifications affect participants' understanding of the app's data practices, privacy concerns, and intentions to grant the permission and download the app. Finally, we compared the impact of justifications contextualized to the app's specific functionalities and how it will use the collected data with non-contextualized generic justifications explaining only what data the app will be able to access.

Overall, we find that adding to the existing permission labels the contextualized justifications about app's data practices improves transparency (in Study 1 and 2) and can mitigate privacy concerns (in Study 2), but does not directly affect the willingness to grant the permissions or download the app. In turn, because privacy concerns negatively affect the intentions to grant permissions, while perceived informativeness of the permissions about app's data practices increases such intentions, eventually, the improved transparency and lower perceived privacy danger increase the willingness to grant permissions. Participants said they generally understand what resources and data will be used by the MAR app based on the given permissions. However, in conditions without justifications, they requested more clarifications about what data is collected by the app, how it is used, whether it is possible to decline the permission, and how it would affect the app's performance. Finally, we find that participants are especially concerned about face and speech recognition, but current systems don't request permissions to run such analysis. Based on the results, we provide recommendations for the improved data transparency in MAR apps' mobile permission systems. Our results hold in the context of data utilized to the users' benefit, and future work is needed to explore other contexts.

2 Related Work

There are two main streams of literature relevant to this work. The first one explores user privacy concerns regarding mobile permissions, and the second one explores user privacy concerns regarding Augmented Reality (AR) technologies.

2.1 Privacy Concerns with Permissions

A plethora of prior research on mobile permissions and privacy-related user perceptions about them confirm that the existing permission systems are not fully transparent and clear to the users [15, 27]. Provision of permissions to users in app stores without any contextual information results in users forgetting about permissions later [6]. Moreover, when apps require more sensitive permissions, it increases users' privacy concerns [18] and decreases the ratings and number of downloads of such apps [30]. Thus, these factors have an immediate economic impact and relevance for app developers and the respective companies [18].

Providing participants with relevant information in runtime permissions is shown to increase transparency for the users [57, 58]. However, besides the general question of when to request permissions [14] and how to simplify them [37], it is still unclear whether the communication of apps' data collection practices in the permission systems can and should be presented in more detail. Particularly, users find it hard to identify the reasons why an app uses a specific resource at all [34]. While requesting app permissions helps users become aware of what data is being accessed by the app, users also want to better understand *why* applications need certain information [27]. By by providing the appropriate justifications and meeting users' expectations regarding the reasons for accessing sensitive resources, apps can increase users' trust [34] and alleviate privacy concerns [18]. However, some research shows that meaningless justifications (that pretend to clarify the purpose of data use, but essentially don't provide any meaningful information) also alleviate user concerns, and therefore can be deceptive for the users [51]. Thus, it is important that permission justifications provide accurate and useful information.

2.2 User Privacy Concerns with AR

Research on privacy in MAR apps is important since context-specific privacy concerns can differ greatly from general privacy concerns towards mobile apps [1, 42]. Although there is a large body of technical research about privacy and security in augmented reality technologies [9], there is little research on end user perceptions and privacy concerns regarding AR technologies, especially, among research focused on *mobile* AR [19]. The limited empirical evidence suggests that AR raises privacy concerns among users, for instance, about being filmed by AR devices (as bystanders [11]), surveillance, and distributing data involuntarily [8, 20, 22, 23, 46].

The analysis of permissions in 19 most downloaded MAR apps in Google Play Store shows that they violate users' privacy and do not follow the principle of least privilege, i.e., apps oftentimes require access to more permissions than they actually need for their stated functionalities [21]. Besides the consumer protection concerns, privacy threats can be a hindering factor for technology adoption [3, 48], which could prevent useful AR applications to be accepted by potential users (e.g., for medical purposes like helping Parkinson's disease patients [36, 54]). Therefore, it is important to understand and address MAR apps' users' privacy concerns.

While prior research has investigated user concerns with mobile permissions and AR technologies separately, to the best of our knowledge no prior work has examined users' privacy concerns regarding the permissions of MAR apps, the impact of permission justifications on these concerns, and intentions to grant permissions to such apps and download them. In this study, we attempt at closing this gap.

3 Study 1

3.1 Method

To answer the research questions, we designed an online survey-based experiment with a between-subject design and three conditions. We presented participants with a scenario describing a fictional mobile augmented reality (MAR) app that can help to redesign a room or outdoor space. We told participants that, by using augmented reality, the app can take and save measurements, or display 3D models of the furniture over the image of the real environment. Also, we told them that users can share the new design ideas and measurements with friends, family, designers, or contractors, via email or in social networks. Then we presented participants with a list of permissions this app requires.

In the *Control* group, participants were presented only with the labels of the permissions without any justifications (e.g. Microphone, Contacts). In the *Contextualized Justification* (CJ) condition, along with the label, we showed participants the explanation of what data or device's sensors will the app access or what data processing approaches will it use (e.g., face or speech recognition) and how it is related to the app's specific functionalities, to help participants understand the purpose of data collection in the context of this particular app. For example, the contextualized justification for the *Microphone* permission mentioned that access to the microphone is required to add voice notes to the measurement photos. In the *Non-Contextualized Justification* (NCJ) condition, the explanation was generic, without adding much to the information in the permission label and without adding context to how the requested permission is related to the app's functionalities and data collection needs. For example, the non-contextualized justification for the *Microphone* permission mentioned that access to the microphone is required to record audio, without explaining why a measurement app would need it. Table 6 in Appendix B provides the text of permission justifications.

We included 7 permissions that are commonly requested in MAR and non-MAR apps: *Storage/Photos/Media Library*, *Contacts*, *Network/Internet Access*, *Microphone*, *Camera*, *Location Services*, and *Notifications*. To account for the customs

of iOS and Android device users, when different, we included labels from both operating systems, e.g., Network/Internet Access. Additionally, we included 9 categories of resources and data processing approaches that are often used in MAR apps, but for which MAR apps currently do not explicitly request user permission (except for *Speech Recognition* on iOS): *Accelerometer*, *Gyroscope*, *Magnetometer*, *LiDAR Scanner*, *Geometry Tracking*, *Raw Camera Output*, *Object Recognition*, *Face Recognition*, and *Speech Recognition*. For simplicity, in this paper we refer to all 16 resources and data processing approaches as *permissions*.

After showing the list of permissions, we asked participants, based on that list, to what extent they understand what data and resources on their device the app will be able to use. We also collected open-ended responses about what additional information would help to improve that understanding. Then we asked whether participants would allow or deny our fictional app access to those permissions on their device, how denying that permission would affect the app's performance, and how granting the permission would affect users' privacy. Finally, we asked about demographics, experience with MAR apps and features, and definition of AR. We also included several attention check questions. See survey in Appendix A.

Quantitative Analysis We used an ordered random-effects logistic regression model to analyze participants' choices regarding granting permissions. The "I am not sure" answers were treated as missing. We calculated the model with three specifications. Model 1 is the base model that includes only the main independent variables about the permissions. Model 2 adds control variables like demographics and AR knowledge to the base model. Model 3 adds further variables based on the five most relevant codes from the qualitative analysis (equals 1 if the participant mentioned the code) to model 2.

We used Shapiro-Wilk tests to assess the normality of data distribution, Wilcoxon rank sum tests for pairwise comparisons, and ANOVA and Kruskal-Wallis equality-of-populations rank tests to assess the differences between the treatment groups. We applied Holm's and Hochberg corrections to all pairwise statistical tests and regressions, and report only the significant results.

Qualitative Analysis To analyze the open-text survey responses we used thematic analysis. Two coders independently developed initial codebooks, merged them, discussed and agreed on the final codebook (Appendix D). They independently applied the codes to all the responses, allowing for multiple attributes per response. Kupper-Hafner interrater agreement rate was 0.84 [31]. Finally, the coders discussed and resolved all the disagreements.

3.2 Participants

We recruited 300 participants using Prolific in June 2020. We restricted participation to US residents, over 18 years old, who use mobile devices on a regular basis, and have approval rates on Prolific over 95%. We excluded 6 responses in which participants failed the attention checks, and 2 responses that were fully identical. The resulting sample consists of 292 participants, which are randomly distributed among three groups: Control ($N = 96$), CJ group ($N = 104$) and NCJ group ($N = 92$). Our sample is sufficient, as power analysis suggested to recruit 85 participants per group to achieve 90% power, with 5% error rate and 0.5 effect size.

The resulting sample has diverse demographics. The participants are 18-74 years old ($mean = 29, SD = 11.40$), 48.63% female and 2.4% prefer to self-identify their gender. About 34% have Bachelor's degree, 31% have done some college but no degree, and 14% have only finished high school; and 31.51% of the participants reported to have a technical background in computer science. ANOVA test confirms no difference in age, gender, and education among the three groups.

Slightly more than half (57.19%) of the participants use an iPhone, and the rest use Android smartphones. The majority of participants choose the correct definition of AR (74.66%) and have experienced AR features (79.45%) like photo masks (e.g., bunny ears in messaging apps) or placing digital objects in the real environment (e.g., AR furniture apps).

3.3 Results

The majority of participants (86.30%) agreed that based on the provided list of permissions they understand what resources and data the app will be able to use (Q3 in Appendix A). On average, compared to the Control group, participants in the CJ group expressed better understanding of what functionalities and data on their device the app will be able to use based on the list of permissions (Wilcoxon rank sum test: $p = 0.0012$). However, there was no difference between NCJ and Control, and between CJ and Non-CJ groups. This means that providing contextualized justification about why the app requires a certain permission and what data it will use significantly improves users' understanding of the app's data practices compared to showing just the permission labels. In contrast, providing the explanations that are not put in context of the specific app do not yield better users' understanding of the app's data practices compared to using just permission labels.

To understand the relative impact of different factors on users' intentions to grant permissions, we conducted regression analysis (Table 7 in Appendix C). Despite significant impact on the ability to understand app's practices (based on the test results), after controlling for other effects, we find no significant treatment effects of justifications on the willingness to grant the permissions. In other words, on average, providing justifications, contextualized or not, does not affect

participants' willingness to grant the permissions when other factors are taken in consideration.

On the other hand, participants are 43% less likely to allow the permission when they believe it will negatively affect app's performance or ability to function (Q8), and 33% less likely to grant the permission when they are concerned about the privacy implications of granting the permission (Q9). In contrast, when participants find the permissions informative, i.e., helpful in understanding what resources and data the app will be able to access (Q11), the odds of granting access to the permissions are 1.245 times higher, given all other variables are held constant.

Other controls like the prior use of AR features, familiarity with the definition of AR, gender, age, education, prior technical experience, smartphone use frequency and mobile OS do not have an effect.

3.3.1 Analysis of Individual Permissions

Willingness to grant the permissions Most participants are willing to allow the permissions (Q7) while the app is in foreground and only few participants would allow the permissions at all times (Figure 1). The majority of participants prefers to deny such permissions as *Contacts*, *Microphone* and *Face Recognition*.

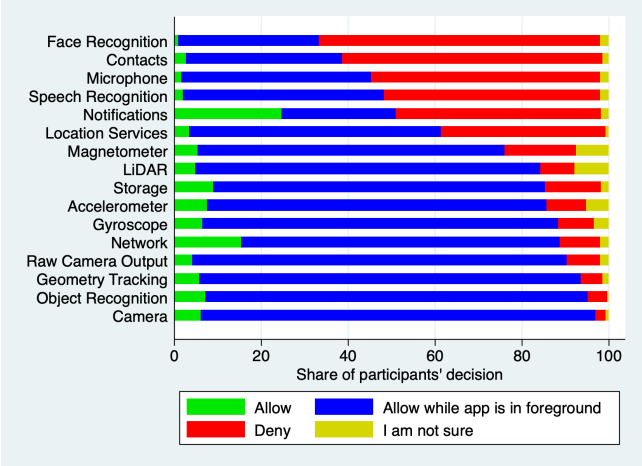


Figure 1: Willingness to grant the permissions (Study 1).

Based on the regression results, we estimated the probabilities for each individual permission to be granted (Table 1). Participants prefer to allow most permissions while in the foreground. However, the likelihood of denying or allowing while in foreground is almost equally split for the following permissions: *Contacts*, *Microphone*, *Location*, *Face Recognition* and *Speech Recognition*.

We also conducted 16 regressions for each of the 16 permissions with the demographic and control variables (same regression model as in Table 7). We did not find any interesting patterns in the results of the individual regressions.

Table 1: Estimated probabilities to deny, allow while in foreground, and allow at all times the permissions (Study 1).

Permissions	Pr_{deny}	$Pr_{foreground}$	Pr_{always}
Storage/Photos/Media Lib.	.239	**.707**	.054
Contacts	**.453**	**.530**	.017
Network/Internet Access	.195	**.728**	.077
Microphone	**.451**	**.533**	.016
Camera	.105	**.772**	.123
Location	**.403**	**.575**	.022
Notifications	.295	**.669**	.036
Accelerometer	.161	**.739**	.100
Gyroscope	.131	**.746**	.123
Magnetometer	.199	**.733**	.068
LiDAR Scanner	.144	**.755**	.101
Geometry Tracking	.111	**.771**	.118
Raw Camera Output	.166	**.750**	.084
Object Recognition	.128	**.759**	.113
Face Recognition	**.506**	**.481**	.013
Speech Recognition	**.433**	**.548**	.019
Total	.257	**.675**	.068

Perceived privacy implications of the permissions To get detailed insights on a permission-specific level, we plot the perceived privacy dangerousness (Q9) of each permission in Figure 2. Participants believe that permissions allowing access to *Face Recognition*, *Contacts*, *Location*, *Microphone*, *Storage* and *Speech Recognition* have the biggest negative impact on their privacy. In contrast, *Magnetometer*, *Accelerometer* and *Gyroscope* are perceived as least privacy invasive.

Participants in the CJ group have, on average, the highest privacy concerns regarding the permissions (mean=4.25 out of 7), followed by the Control group (mean=4.18) and the NCJ group (mean=3.96). The differences between CJ and NCJ as well as Control and NCJ are statistically significant (Wilcoxon rank sum tests: $p = 0.0001$ and $p = 0.0029$, respectively). The difference in the perceived privacy dangerousness between the Control and CJ group is not significant.

Perceived impact on app's performance Overall, participants believe that *denying* the permission would not drastically affect the way the app functions (Q8) (mean=4.03 out of 7). Participants in the NCJ group expected the larger decrease in the app's performance if they deny the permissions, compared to the CJ group (Wilcoxon rank sum tests: $p = 0.0393$; means are 3.95 and 4.12, respectively).

Regarding the effect of *granting* permissions on the device's normal operations (Q10), participants perceive that there is no such negative effect (mean=2.93). There are statistically significant differences between the CJ (mean=3.06) and NCJ (mean=2.72) groups ($p < 0.0001$) and between the Control (mean=2.99) and NCJ groups ($p < 0.0001$).

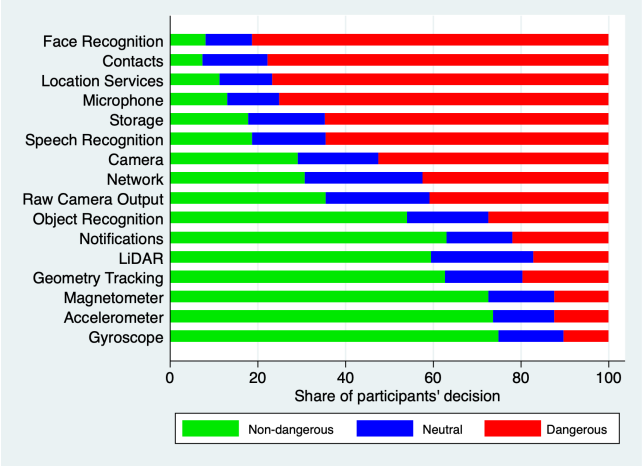


Figure 2: Privacy concerns regarding permissions (Study 1).

Perceived informativeness of the permissions Overall, participants said that they understand what resources and data the app will be able to access if they grant the permissions (Q11) (mean=5.23 out of 7). For example, *Magnetometer* and *LiDAR* are perceived as least informative (means are 3.95 and 4.32, respectively). For all other permissions the means range from 4.95 (*Geometry Tracking*) to 6.22 (*Camera*).

There are statistically significant differences in the participants' perceptions of informativeness (i.e., helpfulness of individual permissions in understanding app's data practices) between the Control (mean=4.78), CJ (mean=5.64) and NCJ (mean=5.25) groups (Wilcoxon rank sum tests $p < 0.0001$ for all three comparisons).

Then, we evaluated the informativeness of the added justifications (compared to solely labels provided in the Control group) according to the following two criteria: 1a) Contextualized justifications should be perceived statistically significantly more informative than the labels alone in the Control group, and 1b) non-contextualized justifications should not. 2) Contextualized justifications should be perceived statistically significantly more informative than non-contextualized ones. We compare the CJ and NCJ group because prior work suggests that practically "meaningless" justifications for the permissions that do not add clarity still may alleviate user concerns [51] as they create a false sense of legitimacy.

We identified that all permissions met Criterion 1, but several permissions did not fulfill Criterion 2. Specifically, our manipulation in the CJ group was effective: compared to the Control group, adding contextualized justifications increases their perceived informativeness, or helpfulness in understanding app's data practices, while non-contextualized justifications do not. Thus, our study provides contrasting evidence compared to the prior work on permission justifications (e.g. [51]).

Criterion 2 is only met by several contextualized justifications: *Face Recognition*, *Speech Recognition*, *Contacts*, *Microphone*, *Storage*, *Network*, and *Accelerometer*. We hypothesized that linguistic complexity [50] might potentially explain that result: some of our contextualized justifications were relatively long and used complex terms in order to provide an informative explanation for the purposes of permission requests.

Thus, it is possible that while contextualized justifications provided more information, this information was harder for the participants to understand than non-contextualized justifications, leading to their reduced perception of informativeness. To address this, in Study 2 we modified the wording of justifications to ensure the equal linguistic complexity so that contextualized and non-contextualized justifications are similar in the required grade level and reading skills to understand them (see Section 4).

3.3.2 Qualitative Results

We asked participants what additional information would help them understand what resources and data on their devices the app will be able to use (Q4). Appendix D provides the codebook, and Table 2 summarizes the most common themes identified in the qualitative analysis of responses. Many participants (74/292), especially in the treatment groups where justifications were provided, said that they do not need any additional information as permission descriptions were already clear enough. However, many other participants said they would like to know more about why the permission is needed and how the data is going to be used (86/292), or requested general clarifications about sensors and features (42/292). Some participants specifically mentioned that they would like the clarifications be concise (17/292), or recommended improving visual representations of the permissions and justifications (11/292), for example, by using demos, screenshots, images, expandable explanations, or grouping the information by topic. Some participants would like to know whether it is possible to deny or restrict individual permissions (18/292) and when the specific data is collected and accessed (13/292).

ANOVA test results indicate that more participants said they do not need additional information in the CJ group ($p < 0.001$) and NCJ group ($p = 0.007$) compared to the Control group. Similarly, the clarifications about sensors or resources were significantly less often requested in the CJ group ($p = 0.023$) and Non-CJ group ($p = 0.031$) than in the Control group. Participants in the CJ group were more often interested to know whether they can deny individual permissions than people in the Control and NCJ group ($p < 0.001$). Participants in the Control group requested information about the purpose of data collection more often than in the CJ ($p < 0.001$) and NCJ group ($p = 0.002$), and they requested this information in NCJ group more often than in CJ group. This result further supports the informativeness of permission justifications, especially the contextualized ones.

In summary, we observed that contextualized justifications

Table 2: Common themes about the additional information needed in the permission justifications, and the ANOVA results of differences between three groups (Study 1, $N = 292$).

Code	Freq., %	ANOVA
Why/how data is used	29.45	F(2)=22.44, p < 0.001
No information needed	25.34	F(2)=8.35, p < 0.001
N/a	15.75	-
Clarifications needed	14.38	F(2)=3.30, p < 0.05
Possibility to deny/ restrict permissions	6.16	F(2)=7.97, p < 0.001
Should be brief / shorter	5.82	F(2)=8.13, p < 0.001
When data is collected/accessed	4.45	-
Visual	3.77	-
Frequency of specifically mentioned permissions		
Microphone	6.51	-
Face recognition	6.51	-
LiDAR	6.16	-
Contacts	5.48	-
Location	3.42	-
Speech recognition	3.42	-

improve the informativeness of the permissions about app's data practices, but do not affect participants' privacy concerns or willingness to grant the permissions, compared to the Control group. However, we observed the need to modify the wording of our justifications to ensure similar linguistic complexity between treatment groups. We also were curious if permissions affect the willingness to download the app. Thus, in Study 2 we modified the wording of our justifications, and added a question about the intention to download the MAR app.

4 Study 2

4.1 Method

The design of Study 2 and its survey (Appendix A) was the same as in Study 1, except several changes. First, to rule out the potential confounding effects due to the difficulty in understanding the justifications' wording, we assessed the linguistic complexity of the contextualized and non-contextualized justifications with the Python library Textstat [45]. This library offers the possibility to measure seven different metrics of a text's readability and complexity levels, and obtain an overall readability score, which combines all seven metrics in one to provide an estimated school grade level required to understand a given text. We used this combined measure to evaluate the complexity of our justifications as it adjusts for biases from single measures such as Flesch-Kincaid or Gunning Fog [45, 49, 55]. Based on the assessment of linguistic complexity, we slightly modified the wording of justifications to simplify and make them similarly easy to understand in the CJ and NCJ conditions (see Table 3).

Second, we added a question about the willingness to download our hypothetical app (Q20) and evaluated factors influencing this download intent using an ordered logistic regression. Finally, we excluded the open-response questions about the participants' suggestions for improving the permissions (Q4-6) as we have already gained enough insights in Study 1 and wanted to keep the survey short.

4.2 Participants

We recruited 306 participants using Prolific in October 2020. We restricted participation to US residents, over 18 years old, who use mobile devices on a regular basis, have approval rates on Prolific over 95%, and have not participated in Study 1. We excluded 16 responses in which participants failed the attention checks, and 1 participant who reported to use a smartphone only about once a year. The resulting sample consists of 289 participants, which are randomly distributed among three groups: Control ($N = 95$), Contextualized Justifications (CJ) ($N = 99$) and Non-Contextualized Justifications (NCJ) ($N = 95$).

The sample composition is similar to Study 1. The participants are 18-69 years old (mean=28.53, SD= 9.59), 57.09% female and 3.81% prefer to self-identify their gender. About a third (30.45%) have a Bachelor's degree, 31.14% have done some college but no degree, and 15.57% have only finished high school; and 29.07% of the participants reported to have a technical background in computer science. ANOVA test confirms that there is no difference in age, gender, and education among the three experimental groups.

Slightly over half of the participants (53%) use Android smartphones, and the rest use Apple's iPhones. The majority of participants choose the correct definition of AR (75.43%) and have experienced AR features (76.12%) like photo masks (e.g., bunny ears in messaging apps) or placing digital objects in the real environment (e.g., AR furniture apps).

4.3 Results

The majority of participants (86.16%) agreed that based on the provided list of permissions they understand what resources and data the app will be able to use (Q3, Appendix A). Participants in the CJ group expressed better understanding than participants in the Control group (Wilcoxon rank sum test: $p = 0.0165$), while there was no significant difference between the NCJ and Control groups, and between the CJ and NCJ groups. These results confirm our findings in Study 1.

Table 3: Permission labels and justifications in Study 2.

Label	Non-Contextualized justification	Contextualized justification
Storage / Photos / Media Library	Access to the smartphone's storage is required to store data processed by the app.	Access to the smartphone's storage is required to save, check, and delete your measurements or furniture ideas.
Contacts	Access to the contacts is required to reach out to your contacts.	Access to the contacts is required to share measurements with your contacts.
Network / Internet Access	Internet access is required to connect the app with the Internet.	Internet access is required to download the images of furniture.
Microphone	Access to the microphone is required to record audio.	Access to the microphone is required to add voice notes to your measurement photos.
Camera	Access to the camera is required to take pictures and videos.	Access to the camera is required to take pictures and videos of the room you are measuring and show the furniture ideas in it.
Location services	Access to location is required to find out where you are.	Access to location is required to filter furniture ideas for those that deliver to your area.
Notifications	Access to notifications is required to send you notifications.	Access to notifications is required to inform you about contacts' comments on furniture ideas.
Accelerometer	Access to the accelerometer is required to improve image stabilization.	Access to the accelerometer is required to improve the image quality and measurements of your room.
Gyroscope	Access to the gyroscope is required to improve image stabilization.	Access to the gyroscope is required to improve the image quality and measurements of your room.
Magnetometer	Access to the magnetometer is required to improve image stabilization.	Access to the magnetometer is required to improve the image quality and measurements of your room.
LiDAR Scanner	Access to the LiDAR scanner is required to illuminate the target with laser light and measure the reflection with a sensor.	Access to the LiDAR scanner is required to measure your room more accurately in low light.
Geometry Tracking	Allowing the app to use geometry tracking is required to create a schematic outline of the environment.	Allowing the app to use geometry tracking is required to measure a schematic outline of a room instead of a real image of it.
Raw Camera Output	Allowing the app to use raw camera output is required to gather and process the real images captured by the camera.	Allowing the app to use raw camera output is required to measure a real image of a room instead of a schematic outline of it.
Object Recognition	Allowing the app to use object recognition is required to detect physical objects.	Allowing the app to use object recognition is required to check if the new furniture (e.g. chairs) would fit with the existing furniture (e.g. table).
Face Recognition	Allowing the app to use face recognition is required to detect faces.	Allowing the app to use face recognition is required for you to log into the app without a password.
Speech Recognition	Allowing the app to use speech recognition is required to identify words and phrases in spoken language.	Allowing the app to use speech recognition is required to turn your voice notes about furniture ideas into text.

The regression analysis (Table 8 in Appendix C) confirms the results from Study 1 as well. We find no treatment effects. Based on the calculations of odds ratios, we find that participants are 40% less likely to grant the permissions when they believe it will negatively affect the app's performance and ability to function (Q8) and 33% less likely when they are concerned about the impact of granting the permissions on their privacy (Q9). Furthermore, they are 1.15 times more likely to grant the permissions when they find them informative (Q11). Moreover, frequent use of the smartphone is negatively associated with the intention to grant the permission (odds ratio = 0.29). Other controls do not have an effect.

4.3.1 Analysis of Individual Permissions

Willingness to grant the permissions As in Study 1, most participants are willing to allow the permissions (Q7) while the app is in foreground and only few participants would allow the permissions at all times (Figure 3). The majority of participants are willing to deny such permissions as *Contacts*, *Microphone*, *Location*, *Face Recognition* and *Speech Recognition* (Table 4). We also find statistically significant negative effects of perceived privacy dangerousness on the intention to grant permissions for several permissions: *Location*, *Contacts*, *Microphone*, *Storage*, *Face Recognition* and *Speech Recognition*, and *Raw Camera Output*. This is not surprising as most of these permissions are perceived especially invasive (Figure 4).

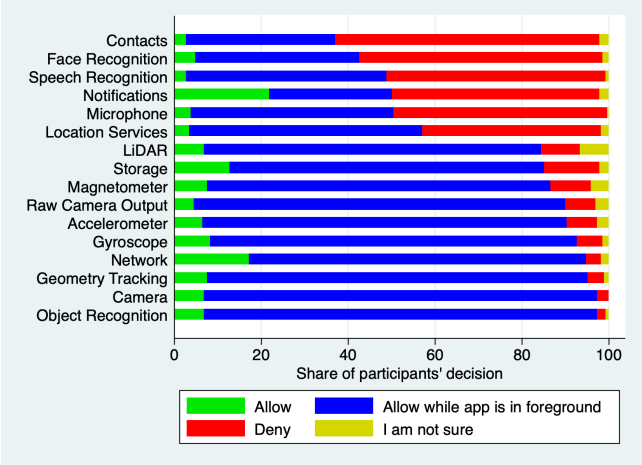


Figure 3: Willingness to grant the permissions (Study 2).

Table 4: Estimated probabilities to deny, allow while in foreground, and allow at all times the permissions (Study 2).

Permissions	Pr_{deny}	$Pr_{foreground}$	Pr_{always}
Storage/Photos/Media Lib.	.229	**.701**	.070
Contacts	**.445**	.534	.021
Network/Internet access	.169	**.732**	.099
Microphone	**.411**	**.562**	.027
Camera	.114	**.754**	.132
Location	**.407**	**.564**	.029
Notifications	.272	**.682**	.046
Accelerometer	.146	**.738**	.116
Gyroscope	.126	**.737**	.137
Magnetometer	.166	**.734**	.100
LiDAR Scanner	.153	**.738**	.109
Geometry Tracking	.120	**.740**	.140
Raw Camera Output	.156	**.743**	.101
Object Recognition	.119	**.747**	.134
Face Recognition	**.438**	**.536**	.026
Speech Recognition	**.408**	**.562**	.030
Total	.242	**.675**	.082

Perceived privacy implications of the permissions Overall, participants believe that the permissions have a moderate effect on their privacy (Q9) (mean=3.94 out of 7). In contrast to Study 1, participants in the CJ group have on average the lowest privacy concerns (mean=3.75), followed by the NCJ group (mean=3.99) and the Control group (mean=4.095). The differences between CJ and NCJ ($p = 0.0031$) as well as Control and CJ ($p < 0.0001$) are statistically significant, while the difference between the Control and NCJ group is not. In other words, while privacy perceptions elicited in the NCJ and Control groups are similar between Study 1 and 2, the modified contextualized justifications in Study 2 elicited less privacy concerns than in the CJ group in Study 1 and than in NCJ and Control groups in Study 2. This suggests that the privacy perceptions regarding permissions can be sensitive to the wording of contextualized justifications.

In line with Study 1, Figure 4 shows that participants perceive permissions allowing access to *Location Services*, *Contacts*, *Face Recognition*, *Microphone*, *Speech Recognition* and *Storage* to have the biggest negative impact on their privacy. *Magnetometer*, *Accelerometer*, *Gyroscope* and *Notifications* are perceived as least privacy invasive. We discuss whether our findings match the categorization of the Android Developer Guide into *normal* and *dangerous* permissions in Section 5.

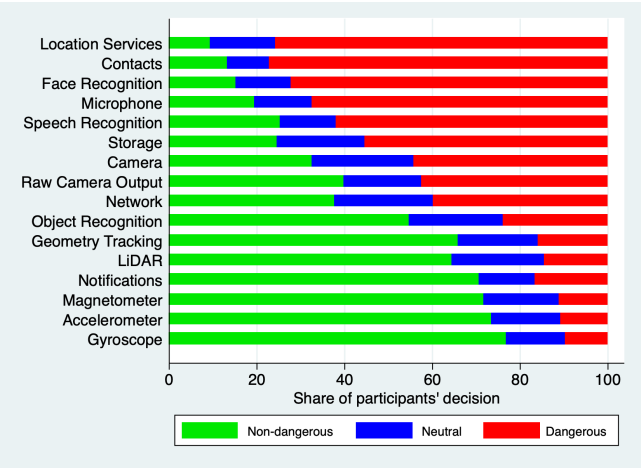


Figure 4: Privacy concerns about the permissions (Study 2).

Perceived impact on app's performance Overall, participants believe that *denying* the permission would not drastically affect the way the app functions (Q8) (mean=3.94 out of 7). As in Study 1, after denying the permission, participants expect the app to function better in the CJ group (mean=4.13) than in the NCJ group (mean=3.82, $p = 0.0001$) and the Control group (mean=3.86, $p = 0.0014$).

Participants perceive that there is no negative effect (mean=2.75) of *granting* permissions on the device's normal operations (Q10). In contrast to Study 1, participants in the CJ condition expected the least negative effect (mean=2.58) compared to the NCJ group (mean=2.7, $p = 0.0492$) and the Control group (mean=2.97, $p < 0.0001$). The difference between the Control and NCJ groups is also significant ($p < 0.0001$). The difference with Study 1 is likely related to the modifications in the wording of justifications in Study 2.

Perceived informativeness of the permissions Overall, participants said that they understand what resources and data the app will be able to access if they grant the permissions (Q11) (mean=5.35 out of 7). *Magnetometer* and *LiDAR* are perceived as least informative (means are 4.22 and 4.49, respectively). For all other permissions the means range from 4.82 (*Accelerometer*) to 6.26 (*Camera*).

We also evaluated to what extent participants' perceptions of informativeness (i.e., helpfulness of individual permissions in understanding app's data practices) differ between treatment and control groups. In line with Study 1, there are statistically significant differences between the Control (mean=4.93), CJ (mean=5.74) and NCJ (mean=5.37) groups (Wilcoxon rank sum tests are $p < 0.0001$ for all three comparisons).

In the next step, we evaluate whether our changes in the wording of justifications based on the insights from Study 1 resulted in any changes in the perceived informativeness of the *individual* permissions. As in Study 1, all permissions with contextualized justifications are perceived as significantly more informative than using only labels (Control group) and there are no statistically significant differences between permissions with non-contextualized justifications and labels only. Contextualized justifications are perceived as significantly more informative than non-contextualized ones only for several permissions: *Network*, *Microphone*, and *Notifications* (Kruskal-Wallis test, p<0.05).

4.3.2 Analysis of Willingness to Download the App

Slightly more than a third of participants (37.72%) said they would be willing to download the app based on the list of permissions. As we assessed the impact of the entire list of permissions, instead of the individual permissions, on the willingness to download the app, there was only one response per participant (i.e. it is not a permission-specific dependent variable). Thus, we could not use the panel structure of the data as in the regressions on the willingness to grant the permissions. Therefore, we created indices for the permission-specific categorical variables Q8-Q11, to estimate the participants' overall perspectives on those variables across all permissions. We calculated a polychoric correlation matrix for each variable and predicted the number of factors for each variable based on the eigenvalues larger than 1, similarly to the procedures of the exploratory factor analysis. The resulting indices and values for Cronbach's α ($0.80 - 0.94$) are shown in Table 5. We then included those indices in the ordered logistic regression models on the willingness to download the app as the dependent variable (see Table 9 in Appendix C).

The only significant variable after the Holm's and Hochberg corrections is the general understanding of the app's data practices. It has a significant positive effect, increasing the intentions to download the app by 1.54 times.

5 Discussion and Conclusions

Overall, prior research suggests that justifications are useful in helping users understand permissions in regular mobile apps [18, 27, 34, 51]. While our work confirms that it is also useful in the MAR apps, it illustrates that justifications are not equally useful, and that they are most effective when explained in the context of a particular app. In addition to the interest in knowing the purpose of data collection, also found in [27], our participants wanted to know which permissions they can restrict/deny and how it would affect the functionality. In contrast to [51], our study finds that non-contextualised justifications do not significantly improve users' understanding of the app's practices. Therefore, justifications should be meaningful and tailored to the context of a particular app to be truly helpful and increase transparency. The share of our participants who chose to deny permissions were similar to prior work conducted in the wild [57, 58], confirming that the intentions observed in our study are likely to be representative of actual users' decisions (however, future work is needed to validate it). Moreover, we explore users' opinions about permissions that are currently not requested in either regular or MAR apps, yet raise significant concerns (e.g. face and speech recognition). In conclusion, our findings have important contributions and practical implications for MAR app developers and permission system design.

5.1 Users' Understanding of and Expectations about Permissions

With respect to our first research question (RQ1), we found that, overall, based on the list of permissions, participants were confident that they understand what resources and data the MAR app will be able to access. However, participants expressed the willingness to know why the MAR app requires certain permissions, and how it is going to use them. We also find that contextualized justifications can increase users' understanding of the app's data practices compared to using just permission labels and could be a useful tool for increasing app transparency. This is especially true for the new permissions that we suggest to add to cover advanced sensors, resources, and functionalities especially common in MAR apps, currently absent in mobile permission systems (e.g., *Magnetometer*, *LiDAR*, *Geometry Tracking* or *Object Recognition*). In other words, adding contextualized justifications to these permissions resulted in the biggest increase in participants' understanding of the app's data practices, compared to the group where only permission labels were used. Open-ended responses confirmed that participants expressed the need for more clarifications about these advanced or novel sensors and data processing approaches (especially *Magnetometer*, *Geometry Tracking* and *LiDAR*).

Participants believed that refusing to grant certain non-essential permissions (e.g., send notifications) would not impair the app's ability to perform its primary functionalities. In the open-text responses, participants also said that they would like to know whether they can decline some of the permissions. Thus, we recommend to have a clear labeling of which permissions are required and which permissions are optional, and how declining of such permissions would affect the app's functionalities (e.g., if users were to decline the *Notifications* permissions, they would not be able to receive

Table 5: Indices created from variables Q8-Q11 and Cronbach's α

Variable	Index	Permissions	Alpha
Q8: App functioning w/o accessing permission X	1	Storage, network, camera, accelerometer, gyroscope, magnetometer, LiDAR, geometry tracking, raw camera tracking, object recognition	0.8442
	2	Contacts, microphone, location, notifications, face recognition, speech recognition	0.7977
Q9: Privacy dangerousness of permission X	1	Notifications, accelerometer, gyroscope, magnetometer, LiDAR, geometry tracking, raw camera output, object recognition	0.8890
	2	Storage, contacts, network, microphone, camera, location, face recognition, speech recognition	0.8852
Q10: Negative effects on performance of permission X	1	All 16 permissions	0.9438
Q11: Perceived informativeness of permission X	1	Storage, contacts, network, microphone, camera, location, notifications, face recognition, speech recognition	0.8569
	2	Accelerometer, gyroscope, magnetometer, LiDAR, geometry tracking, raw camera output, object recognition	0.8934

notifications about the comments that the designer has left about their measurements in the MAR app).

Similarly, participants did not believe that granting the permissions could negatively affect the device's normal operations. This question was inspired by the Android Developer Guide, which classifies dangerous permissions as those that have an impact on user's privacy and the device's normal operations [2]. As we discuss in Section 5.2, participants expressed privacy concerns about certain permissions, but did not expect a negative impact on device's normal operations. However, it is possible that the assessment of the impact on the device's normal operations may require more advanced technical knowledge than the assessment of privacy implications. Without proper guidance on how to evaluate the impact of permissions on the device's normal operations, it would be hard to decide whether the new permissions, such as *LiDAR* and *Geometry Tracking*, should be categorized as dangerous or not. Thus, we recommend the Android Developer Guide (and other similar documentations) to include more details about the metrics used to make such assessments, and how to reconcile the contradictions between negative impact on user's privacy and no impact on device's normal operations.

5.2 Privacy Concerns about Permissions

Regarding our second research question (RQ2), some permissions (e.g., *Location*, *Contacts*, *Microphone*, *Speech and Face Recognition*) raise substantial privacy concerns among users. Prior work found that users are concerned about access to location and microphone in non-AR apps as well [15, 39, 59]. However, most apps do not request the permissions to use *Speech and Face Recognition*, although users in our study find it concerning. It is possible that participants are especially concerned about *Speech and Face Recognition* due to the general lack of understanding of what information is collected for it, at what point in time, and how it is processed. Media coverage of the privacy invasions by companies relying on face recognition technologies such as Clearview AI [25] could also trigger privacy concerns among our participants.

In contrast, other permissions, such as *Accelerometer*, *Magnetometer*, *Gyroscope*, *Notifications*, *LiDAR*, and *Geometry Tracking*, did not raise high privacy concerns. However, it has been shown that accelerometer data, which is collected by several MAR apps like Pokémon Go, can be used to infer the phone's password [16] or to eavesdrop on the audio output of the device [35]. This findings indicates a gap between users' understanding of the threat models and actual privacy and security risks. This kind of knowledge-concern gap is critical for MAR apps. The variety of sensors required by MAR apps and the technical possibilities to exploit these sensors make it important to inform the users about the potential privacy implications, and request the permissions to access those resources. Clarifications and justifications for the sensors could further help to bridge this gap. The accuracy and informativeness of these justifications should be enforced and monitored, for example, by the app stores and regulators.

We also checked if users' privacy perceptions align with the Android Developer Guide classification of permission dangerousness [2]. We only considered the responses of participants in the Control groups (in both Studies 1 and 2), as currently the permission systems use only the labels, without justifications. We categorized participants' responses based on the median values for the individual permissions into two groups: (1) Dangerous permissions, if they are perceived on average as privacy invasive (median value larger than 4 out of 7), and (2) Non-Dangerous permissions, if they are perceived on average as not privacy invasive or neutral (median value less than or equal to 4 out of 7). The results indicate that the Android's classifications match the participants' evaluations for the currently used permissions. However, it also suggests that if new permissions were added, some of them (such as *Face and Speech Recognition*) would be considered dangerous and would need to be requested in the run time.

5.3 The Impact of Justifications

Regarding the third research question (RQ3), our results indicate that contextualized justifications, which describe what information the app will be able to access and how it will use it if the user grants the permission, improve users' understanding of the app's data practices, and reduce their privacy concerns, but do not impact users' intentions to grant such permissions or download the app. In turn, privacy concerns reduce the intentions to grant the permissions, but not to download the app. This might be due to the fact that users' decisions to download an app are more dependent on other factors, such as app's functionalities and utility. We find that main results are similar in Study 1 and 2, indicating the robustness of the effects against the modifications in the wording of the permission justifications. Thus, we recommend app developers and platforms to include contextualized justifications to the existing permission systems to improve the transparency about MAR apps' data practices. As discussed in Section 5.5, while our study explores the opinions about only one type of MAR apps, future work is encouraged to validate the results with other categories of MAR apps, and different levels of invasiveness of their data practices.

5.4 Suggestions for Improved Transparency

Regarding our fourth research questions (RQ4), we provide a number of recommendations for improved transparency in the permission systems of MAR apps based on all our results. First, we recommend app platforms to require developers to provide justifications for the permissions requested by their apps. Second, we encourage app developers to contextualize those justifications to the specific practices and functionalities of their apps. For example, instead of using vague statements about the app's need to access the *Microphone* in order to transmit audio input, we suggest explaining how it will be used by the app (e.g., to record voice messages).

Third, we recommend requesting user permissions to access the resources and sensors that are commonly used in MAR apps and often raise privacy concerns, but are not currently included in the mobile permission systems, such as *Face and Speech Recognition*. Similarly, as technology advances very fast, we recommend including a short description of the novel functionalities and sensors, such as *LiDAR*, *Geometry Tracking* and *Object Tracking*, avoiding technical terminology that can be hard to understand for the people with limited technological background or experience. We also recommend testing the linguistic complexity of the justifications (e.g., by using the library we used in this study [45]), and the overall comprehension and informativeness of the justifications in user studies.

Fourth, based on the participants' comments, we recommend improving the visual appearance of the permission systems. For instance, we suggest:

- Group the permissions by the type of information they access or by the purpose of use;
- Avoid permission fatigue and allow customisation of the level of detail in permission justifications, for example, by using expandable clarifications text boxes, larger or higher-contrast text for labels and less prominent justification text to allow users to quickly scan the permissions and easily read the additional information where they need more clarifications;
- Include images, videos, or animations to demonstrate how advanced or novel sensors work;
- Make it clear *when* certain functionalities, sensors or data are being accessed;
- Clearly indicate whether a certain permission can be denied or restricted, and how the restrictions of certain permissions can affect the performance of the app;
- Clarify privacy implications of the permissions.

5.5 Limitations and Future Work

Our study has several limitations. First, while our sample is diverse in terms of demographics, it only includes US citizens. Incorporating cultural factors can provide additional insights for privacy-related predictions [33, 53]. In the future work, we would like to expand the diversity of the sample and conduct a cross-country comparison of users' perceptions and understanding of MAR apps' mobile permissions.

Second, to keep high internal validity, we tested only one treatment dimension related to the permission justifications (contextualized and non-contextualized), while keeping other parameters constant. Future work can experiment with other dimensions, such as the number and composition of requested permissions, their relevance or importance to the app's functionalities, purposes of data use (including the purposes that primarily benefit the companies more than users, such as targeted advertising), or visual design of the permission menus.

Finally, we used a hypothetical scenario about a MAR app in our study. However, the use of hypothetical scenario in a controlled experiment allowed us to achieve high internal validity, and future work can test the generalizability and ecological validity of the results in a field experiment. Moreover, the app we used in our scenario can be categorized as a utilitarian app with primarily utilitarian incentives for individuals to use it. In contrast, hedonic (or pleasure and entertainment oriented) MAR apps, such as games, could be judged differently by participants regarding their privacy implications, willingness to grant permissions, or download intentions. Thus, future work can explore the differences in users' opinions and intentions regarding various categories of MAR apps.

Acknowledgments

We thank Julia Bernd, Serge Egelman, other BLUES members, and CLTC Research Symposium participants for their comments on the study design and suggestions about the paper. This work was supported by the Center for Long-Term Cybersecurity at UC Berkeley, and National Science Foundation grants CNS-1514211 and CNS-1528070.

References

[1] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.

[2] Android Developers. Android Permissions Overview. https://developer.android.com/guide/topics/permissions/overview#permission-groups, 2019.

[3] Corey M. Angst and Ritu Agarwal. Adoption of Electronic Health Records in the Presence of Privacy Concerns: The Elaboration Likelihood Model and Individual Persuasion. *MIS Ouarterly*, 33(2):339–370, 2009.

[4] Ronald T. Azuma, Yohan Baillot, Steven Feiner, Simon Julier, Reinhold Behringer, and Blair Macintyre. Recent Advances in Augmented Reality. *IEEE Computer Graphics And Applications*, 21(6):34–47, 2001.

[5] Gökhan Bal, Kai Rannenberg, and Jason I. Hong. Styx: Privacy risk communication for the Android smartphone platform based on apps' data-access behavior patterns. *Computers & Security*, 53(September):187–202, sep 2015.

[6] R. Balebako, F. Schaub, I. Adjerid, A. Acquisti, and L. Cranor. The impact of timing on the salience of smartphone app privacy notices. In *Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices*, pages 63–74, 2015.

[7] Rochelle A Cadogan. An imbalance of power: the readability of internet privacy policies. *Journal of Business & Economics Research (JBER)*, 2(3), 2011.

[8] Scott G. Dacko. Enabling smart retail settings via mobile augmented reality shopping apps. *Technological Forecasting and Social Change*, 124:243–256, 2017.

[9] Jaybie A. de Guzman, Kanchana Thilakarathna, and Aruna Seneviratne. Security and Privacy Approaches in Mixed Reality: A Literature Survey. 2018.

[10] Dejan G. 29+ Augmented Reality Stats to Keep You Sharp in 2020. https://techjury.net/blog/augmented-reality-stats/, 2020.

[11] Tamara Denning, Zakariya Dehlawi, and Tadayoshi Kohno. In situ with bystanders of augmented reality glasses. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*, pages 2377–2386, 2014.

[12] Arindam Dey, Mark Billinghurst, Robert W Lindeman, and J. Edward Swan II. A Systematic Review of Usability Studies in Augmented Reality between 2005 and 2014. In *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, pages 49–50, Merida, 2016.

[13] W. Enck, P. Gilbert, B. Chun, L. P. Cox, J. Jung, P. McDaniel, and A. N. Sheth. Taintdroid: an information-flow tracking system for realtime privacy monitoring on smartphones. In *9th ACM USENIX Conference on Operating Systems Design and Implementation*, pages 393–407, 2010.

[14] Adrienne Porter Felt, Serge Egelman, Matthew Finifter, Devdatta Akhawe, David A Wagner, et al. How to ask for permission. *HotSec*, 12:7–7, 2012.

[15] Adrienne Porter Felt, Elizabeth Ha, Serge Egelman, Ariel Haney, Erika Chin, and David Wagner. Android permissions: User attention, comprehension, and behavior. In *Symposium on Usable Privacy and Security (SOUPS)*, pages 1–14, 2012.

[16] Duncan Geere. Scientists find a way to crack your phone's password using just the accelerometer. https://www.techradar.com/uk/news/scientists-find-a-way-to-crack-your-phones-password-using-\just-the-accelerometer, 2017.

[17] Alessandra Gorla, Ilaria Tavecchia, Florian Gross, and Andreas Zeller. Checking app behavior against app descriptions. In *Proceedings of the 36th international conference on software engineering*, pages 1025–1035, 2014.

[18] Jie Gu, Yunjie (Calvin) Xu, Heng Xu, Cheng Zhang, and Hong Ling. Privacy concerns for mobile app download: An elaboration likelihood model perspective. *Decision Support Systems*, 94:19–28, 2017.

[19] David Harborth. Augmented Reality in Information Systems Research: A Systematic Literature Review. In *Twenty-third Americas Conference on Information Systems (AMCIS)*, pages 1–10, Boston, 2017.

[20] David Harborth. Unfolding Concerns about Augmented Reality Technologies: A Qualitative Analysis of User Perceptions. In *Wirtschaftsinformatik (WI19)*, pages 1262–1276, 2019.

[21] David Harborth, Majid Hatamian, Welderufael B. Tesfay, and Kai Rannenberg. A Two-Pillar Approach to Analyze the Privacy Policies and Resource Access Behaviors of Mobile Augmented Reality Applications. In *Hawaii International Conference on System Sciences (HICSS) Proceedings*, pages 5029–5038, 2019.

[22] David Harborth and Sebastian Pape. Privacy Concerns and Behavior of Pokémon Go Players in Germany. In M. Hansen, E. Kosta, I. Nai-Fovino, and S. Fischer-Hübner, editors, *Privacy and Identity Management. The Smart Revolution. Privacy and Identity 2017. IFIP Advances in Information and Communication Technology, vol 526*, pages 314–329. Springer, Cham, 2018.

[23] David Harborth and Sebastian Pape. Investigating Privacy Concerns related to Mobile Augmented Reality Applications. In *International Conference on Information Systems (ICIS)*, pages 1–9, 2019.

[24] Majid Hatamian, Jetzabel Serna, Kai Rannenberg, and Bodo Igler. FAIR: Fuzzy Alarming Index Rule for Privacy Analysis in Smartphone Apps. In *International Conference On Trust, Privacy & Security In Digital Business (TrustBus 2017)*, pages 1–16, 2017.

[25] Kashmir Hill. The Secretive Company That Might End Privacy as We Know It. https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html, 2020.

[26] Patrick G. Kelley, Sunny Consolvo, Lorrie F. Cranor, Jaeyeon Jung, Norman Sadeh, and David Wetherall. A conundrum of permissions: installing applications on an android smartphone. In *Proceedings of the 26th International Conference on Fin. Cryptography and Data Security*, pages 68–79, 2012.

[27] Patrick Gage Kelley, Lorrie Faith Cranor, and Norman Sadeh. Privacy as part of the app decision-making process. *SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*, pages 3393–3402, 2013.

[28] Vanessa Kohn and David Harborth. Augmented reality – A game changing technology for manufacturing processes? In *Twenty-Sixth European Conference on Information Systems (ECIS2018)*, pages 1–19, Portsmouth, UK, 2018.

[29] Kate Kozuch. Apple Glasses: Release date, price, features and leaks. https://www.tomsguide.com/news/apple-glasses, 2021.

[30] M. Kummer and P. Schulte. When private information settles the bill: Money and privacy in Google's market for smartphone applications. *Management Science*, 65(8):3470–3494, 2019.

[31] Lawrence L Kupper and Kerry B Hafner. On assessing interrater agreement for multiple attribute responses. *Biometrics*, pages 957–967, 1989.

[32] Kiron Lebeck, Kimberly Ruth, Tadayoshi Kohno, and Franziska Roesner. Towards Security and Privacy for Multi-user Augmented Reality: Foundations with End Users. In *2018 IEEE Symposium on Security and Privacy*, pages 392–408, 2018.

[33] Yao Li, Alfred Kobsa, Bart P Knijnenburg, and MH Carolyn Nguyen. Cross-cultural privacy prediction. *Proceedings on Privacy Enhancing Technologies*, 2017(2):113–132, 2017.

[34] J. Lin, S. Amini, J. I. Hong, N. Sadeh, J. Lindqvist, and J. Zhang. Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, pages 501–510, 2012.

[35] Malwarebytes Labs. The little-known ways mobile device sensors can be exploited by cybercriminals. https://blog.malwarebytes.com/iot/2019/12/the-little-known-ways-mobile-device-sensors-can-be-exploited-by-cybercriminals/, 2019.

[36] Rísin McNaney, John Vines, Daniel Roggen, Madeline Balaam, Pengfei Zhang, Ivan Poliakov, and Patrick Olivier. Exploring the Acceptability of Google Glass as an Everyday Assistive Device for People with Parkinson's. In *32nd annual ACM Conference on Human factors in Computing Systems*, pages 2551–2554, 2014.

[37] Kristopher Micinski, Daniel Votipka, Rock Stevens, Nikolaos Kofinas, Michelle L Mazurek, and Jeffrey S Foster. User Interactions and Permission Use on Android. *SIGCHI Conference on Human Factors in Computing Systems (CHI '17)*, pages 362–373, 2017.

[38] Microsoft. Microsoft HoloLens. https://www.microsoft.com/microsoft-hololens/en-us/buy, 2017.

[39] Alexios Mylonas, Marianthi Theoharidou, and Dimitris Gritzalis. Assessing Privacy Risks in Android : A User-Centric Approach. In *International Workshop on Risk Assessment and Risk-driven Testing*, pages 21–37. Springer, Cham, 2013.

[40] Randy Nelson. Pokémon GO Revenue Hits $1.8 Billion on Its Two Year Launch Anniversary. https://sensortower.com/blog/pokemon-go-revenue-year-two, 2018.

[41] Jack Nicas and Cat Zakrzewski. Augmented Reality Gets Boost From Success of 'Pokémon Go'. https://www.wsj.com/articles/augmented-reality-g

ets-boost-from-success-of-pokemon-go-\1468402203, 2016.

[42] Helen Nissenbaum. *Privacy in Context: Technology, Policy and the Integrity of Social Life*. Stanford University Press, Palo Alto, 2010.

[43] Victoria Petrock. US Virtual and Augmented Reality Users 2020. https://www.emarketer.com/content/us-virtual-and-augmented-reality-users-2020, 2020.

[44] Robert W Proctor, M Athar Ali, and Kim-Phuong L Vu. Examining usability of web privacy policies. *Intl. Journal of Human–Computer Interaction*, 24(3):307–328, 2008.

[45] Python Software Foundation. textstat 0.7.0. https://pypi.org/project/textstat/, 2021.

[46] Philipp A. Rauschnabel, Jun He, and Young K. Ro. Antecedents to the adoption of augmented reality smart glasses: A closer look at privacy risks. *Journal of Business Research*, 92:374–384, 2018.

[47] Lisa Rosenfeld, John Torous, and Ipsit V Vahia. Data security and privacy in apps for dementia: an analysis of existing privacy policies. *The American Journal of Geriatric Psychiatry*, 25(8):873–877, 2017.

[48] Craig Van. Slyke, Richard Johnson, James Jiang, and J.T. Shim. Concern for Information Privacy and Online Consumer Purchasing. *Journal of the Association for Information Systems*, 7(6):415–444, 2006.

[49] Ali Sunyaev, Tobias Dehling, Patrick L Taylor, and Kenneth D Mandl. Availability and quality of mobile health app privacy policies. *Journal of the American Medical Informatics Association*, 22(e1):e28–e33, 2015.

[50] Benedikt Szmrecsanyi. An informationtheoretic approach to assess linguistic complexity. *Complexity, isolation, and variation*, 57:71, 2016.

[51] Joshua Tan, Khanh Nguyen, Michael Theodorides, Heidi Negrón-Arroyo, Christopher Thompson, Serge Egelman, and David Wagner. The effect of developer-specified explanations for permission requests on smartphone user behavior. In *SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*, pages 91–100, 2014.

[52] Kaitlyn Tiffany. It's cool to look terrifying on pandemic instagram. https://www.theatlantic.com/technology/archive/2020/05/augmented-reality-instagram-zoom/611494/, 2020.

[53] Sabine Trepte, Leonard Reinecke, Nicole B Ellison, Oliver Quiring, Mike Z Yao, and Marc Ziegele. A cross-cultural perspective on the privacy calculus. *Social Media + Society*, 3(1):2056305116688035, 2017.

[54] Erik van der Meulen, Marina-Anca Cidotã, Stephan G Lukosch, Paulina J M Bank, Aadjan J C van der Helm, and Valentijn T Visch. A Haptic Serious Augmented Reality Game for Motor Assessment of Parkinson's Disease Patients. In Eduardo E Veas, Tobias Langlotz, José Martinez-Carranza, Raphaël Grasset, Maki Sugimoto, and Alejandro Martin, editors, *International Symposium on Mixed and Augmented Reality, ISMAR 2016 Adjunct*, pages 102–104. IEEE, 2016.

[55] Tiffany M Walsh and Teresa A Volsko. Readability assessment of internet-based consumer health information. *Respiratory Care*, 53(10):1310–1315, 2008.

[56] Xuetao Wei, Lorenzo Gomez, Iulian Neamtiu, and Michalis Faloutsos. Permission evolution in the android ecosystem. In *Proceedings of the 28th Annual Computer Security Applications Conference*, pages 31–40, 2012.

[57] Primal Wijesekera, Arjun Baokar, Ashkan Hosseini, Serge Egelman, David Wagner, and Konstantin Beznosov. Android Permissions Remystified: A Field Study on Contextual Integrity. In *Proceedings of the 24th USENIX Security Symposium*, 2015.

[58] Primal Wijesekera, Arjun Baokar, Lynn Tsai, Joel Reardon, Serge Egelman, David Wagner, and Konstantin Beznosov. The feasibility of dynamically granted permissions: Aligning mobile privacy with user preferences. In *2017 IEEE Symposium on Security and Privacy*, pages 1077–1093. IEEE, 2017.

[59] Yixin Zhang and Ryan Gilbert Garcia. Do users really care about privacy? Mobile applications' popularity, user satisfaction, and permission requests. In *Twenty-Eight European Conference on Information Systems (ECIS2020)*, 2020.

All websites were last accessed January 11, 2021.

A Questionnaires

All questions are the same in both studies, except: Q4-6 are included only in Study 1, Q20 are included only in Study 2.

Part I. AR Knowledge

Q1. What is the definition of Augmented Reality? *[if answered incorrectly, participants get the correct definition of AR]* 1. Augmented Reality is the perception of a completely virtual environment in which the user is fully immersed. 2. Augmented Reality is the real environment enhanced by virtual information and objects in which the user is able to perceive the real environment. 3. Augmented Reality combines controlled steering of laser beams with a laser rangefinder in order to measure surfaces or bodies to generate a picture.

Q2. Some mobile applications (apps) have Augmented Reality features, which augment the real environment by virtual information and objects, like photo masks. These features may be required for the app to function (e.g. an AR game which is impossible to play without using AR features), or may be optional (e.g. a photo filter in a messaging app). What Augmented Reality features do you use in the apps installed on your phone? Choose all that apply: 1. Photo masks which add digital objects to the photo (e.g. bunny ears to your face, stars, special effects). 2. Digital representations of objects in real environments (e.g. furniture added into existing view of a room). 3. Displaying digital game characters and game worlds' objects in the real environments (e.g. Pokémon Go's). 4. Other (please specify).

AC1. Please choose the answer option 'always' here. (1. Never. 2. Sometimes. 3. About half of the time. 4. Most of the time. 5. Always.)

Part II. Permission Overview

We would appreciate your feedback on an Augmented Reality app that we are developing for mobile devices. Please read the description of the app carefully before answering the following questions.

The new 'Measure it! Augmented Reality App' allows you to redesign a room or outdoor space. Using Augmented Reality it can take and save measurements, or try out new furniture by displaying its 3D models over the image of the real environment. Plus, with just a few clicks, you can easily share the new design ideas and measurements with friends, family, your designer, or contractors, via email or in social networks! The app requires access to the following functionalities and data on your device: (See the list of permissions in Appendix B and Table 3.)

Q3. Based on the list of permissions above, to what extent do you understand what functionalities and data on your device the app will be able to use? (7pt Likert scale from "I don't understand at all" to "I fully understand")

Q4.* What additional information would help you to understand what functionalities and data on your device the app will be able to use? (open text)

Q5.* What other functionalities of your device do you think the app may be using that are not included in the listed permissions? (open text)

Q6.* What other data do you think the app may be using that are not included in the listed permissions? (open text)

Part III. Evaluating Individual Permissions

AC2. This question is not part of the survey and just helps us to detect bots and automated scripts. To confirm that you are a human, please choose 'strongly agree' here. (7pt Likert scale from "Strongly disagree" to "Strongly agree")

The following questions are iterated for each permission. Imagine that you received the following notification on your phone: "The app Measure it! Augmented Reality needs to access [permission] on your device."

Q7. Would you allow or deny the app to access your device's [permission]? 1. Deny. 2. Allow while app is in foreground. 3. Allow. 4. I'm not sure (if this is selected: "Under what circumstances would you allow or deny this permission?").

Q8. How well do you think the app can function without accessing [permission] on your device? (Consider that 1 star is when the app cannot function without it at all, and 7 starts is when the app can function perfectly without it.)

Q9. To what extent do you think that granting permission to access [permission] on your device can potentially affect your privacy? (7pt Likert scale from "No effect" to "Very big effect")

Q10. To what extent do you think that granting permission to access [permission] on your device can potentially affect your device's normal operations (i.e. performance)? (7pt Likert scale from "No effect" to "Very big effect")

Q11. To what extent do you understand what functionalities and data the app will be able to use, if you allow it to access [permission] on your device? (7pt Likert scale from "I don't understand at all" to "I fully understand")

Q20.* How likely are you to download this app? (7pt Likert scale from "Extremely unlikely" to "Extremely likely").

Part IV. Demographics

Q12. What is your gender? 1. Male. 2. Female. 3. Prefer to self-identify. 4. Prefer not to say.

Q13. What is your age? (numeric entry field)

Q14. What is your country of residence (US or other)

Q15. What is the highest level of school you have completed or the highest degree you have received? 1. Less than high school degree. 2. High school graduate (high school diploma or equivalent including GED). 3. Some college but no degree. 4. Associate degree in college (2-year). 5. Bachelor's degree in college (4-year). 6. Master's degree. 7. Doctoral degree. 8. Professional degree (JD, MD).

Q16. Do you have experience in any of the following (choose all that apply)? 1. Computer science education / work experience2. 2. Software engineering education / work experience. 3. App development education / work experience. 4. Other technical education / work experience (please specify). 5. None of the above.

Q17. How often do you use a smartphone? (from "Never" to "Once or several times a day")

Q18. Which operating system do you use on your smartphone? (Android, iOS, other)

Q19. Do you have any feedback regarding the questionnaire or the study? (open text)

B Permissions and Justifications (Study 1)

Table 6: Permission labels and justifications in Study 1.

Label	Non-Contextualized justification	Contextualized justification
Storage / Photos / Media Library	Access to the smartphone's storage is required to store data processed by the app.	Access to the smartphone's storage is required to browse and edit (save, erase) the photographs of the taken measurements.
Contacts	Access to the contacts is required to enable social features of the app.	Access to the contacts is required to share photos of the measurements with the contacts via email or messages (for example, with your designer, contractors, partner, or friends).
Network / Internet Access	Internet access is required to connect the app with the Internet.	Internet access is required to share your measurement photos via email, messengers, or in social networks (for example, with your designer, clients, contractors, partner, or friends).
Microphone	Access to the microphone is required to record audio.	Access to the microphone is required to add voice notes to your measurement photos.
Camera	Access to the camera is required to take pictures and videos.	Access to the camera is required to take photos and videos of the environments you are measuring. These photos and videos allow the app to visualize your furniture and other objects in those environments.
Location services	Access to location information is required to detect the approximate position (based on network data) and precise position (based on GPS and network data) of the device.	Access to location information is required to link your measurement photos to location data. This information allows the app to automatically create albums in your gallery based on location, which makes it easier to navigate through your measurements.
Notifications	Access to notifications is required to notify you about messages.	Access to notifications is required to notify you when new messages or comments about measurements or furniture ideas are received from your contacts (e.g., designer, clients, contractors, partner, or friends).
Accelerometer	Access to the accelerometer is required to measure the acceleration of your smartphone movements.	Access to the accelerometer is required to provide image stabilization based on the speed your phone is moving, which improves the quality of your measurement photos.
Gyroscope	Access to the gyroscope is required to measure the rotation of the smartphone.	Access to the gyroscope is required to provide image stabilization based on the position of your device and the vibrations of your hands.
Magnetometer	Access to the magnetometer is required to measure magnetic fields.	Access to the magnetometer is required to detect nearby magnetic fields, which can reduce the accuracy and quality of your measurement photos.
LiDAR Scanner	Access to the LiDAR scanner is required to use light to measure distances.	Access to the LiDAR scanner is required to provide detailed 3D measurements of your environment. This allows the app to more accurately represent the location of furniture and other objects.
Geometry Tracking	Allowing the app to use geometry tracking is required to generate a geometrical schematic outline of the environment.	Allowing the app to use geometry tracking is required to generate a geometrical schematic outline of the environment for measuring distances between objects, instead of using the raw camera output of that environment (i.e. real views of the environments, such as rooms, or outdoor spaces and objects in them).
Raw Camera Output	Allowing the app to use the raw camera output is required to gather and process the raw output of the camera while you use the app.	Allowing the app to use the raw camera output is required to present you with a realistic presentation of the pieces of furniture in the real environment, instead of just a geometrical schematic outline.
Object Recognition	Allowing the app to use object recognition is required to recognize details of the objects in the environments.	Allowing the app to use object recognition is required to identify objects in your environment (e.g. the existing furniture in the room). This allows the app to remove or substitute those objects with augmented reality objects, such as viewing how a new couch would fit in the room, or whether new chairs would fit with the existing table.
Face Recognition	Allowing the app to use face recognition is required to identify or verify the identities of people using their face.	Allowing the app to use face recognition is required to verify the identity of people in your measurement photos. This allows the app to automatically identify people in your device's contacts list if they appear in your measurement photos, so that you can easily share among people in the environment for easy sharing of the measurements and furniture ideas.
Speech Recognition	Allowing the app to use speech recognition is required to identify words and phrases in spoken language and convert them to a machine-readable format.	Allowing the app to use speech recognition is required to transcribe voice notes into text for the taken measurements or furniture ideas.

C Regression Analyses, Factor Analysis

Table 7: Random-effects ordered logistic regression models on the willingness to grant permissions in Study 1.

	(1) Base Model	(2) With Controls	(3) With Qual. Vars
Dependent variable: Would you allow or deny the app to access your device's permission X? (Q7)			
Contextualized Justifications experimental group (control group is omitted)	0.190	0.166	0.268
	(0.95)	(0.85)	(1.14)
Non-Contextualized Justifications experimental group	0.175	0.211	0.259
	(0.83)	(1.03)	(1.19)
Q8: App functioning w/o accessing permission X	**-0.567*** **	**-0.567*** **	**-0.566*** **
	(-12.40)	(-12.40)	(-12.38)
Q9: Privacy dangerousness of permission X	**-0.396*** **	**-0.396*** **	**-0.397*** **
	(-12.03)	(-12.01)	(-12.00)
Q10: Negative effects on performance of permission X	0.038	0.041	0.040
	(0.87)	(0.94)	(0.92)
Q11: Perceived informativeness of permission X	**0.218*** **	**0.218*** **	**0.219*** **
	(5.33)	(5.30)	(5.32)
Q3: General app understanding	0.011	-0.006	0.004
	(0.15)	(-0.08)	(0.06)
Q1: Definition of AR correct		-0.120	-0.080
		(-0.66)	(-0.44)
Q2: AR features on smartphones used		0.475*	0.526*
		(2.13)	(2.40)
Q12_1: Gender - male ("female" is omitted)		0.084	0.061
		(0.47)	(0.34)
Q12_2: Gender (prefer to self-identify)		-0.162	-0.193
		(-0.36)	(-0.42)
Q13: Age		0.001	-0.000
		(0.09)	(-0.01)
Q15: Education		-0.119*	-0.118*
		(-2.12)	(-2.13)
Q16: Technically experienced		-0.033	-0.047
		(-0.17)	(-0.24)
Q17: Smartphone used once or several times a day		-0.626	-0.743
		(-0.67)	(-0.81)
Q18: Mobile OS (1=Android)		0.073	0.130
		(0.41)	(0.74)
Q4_1: Why and how data is used			-0.169
			(-0.86)
Q4_2: No information needed / clear what the permission(s) is and does			-0.249
			(-1.07)
Q4_3: Clarification about sensors / features needed			0.112
			(0.57)
Q4_4: Possibility to deny / restrict individual permissions			**-0.828*** **
			(-3.59)
Q4_5: When data is collected			0.257
			(0.72)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; t statistics in parentheses. **Values in bold font indicate statistical significance that hold after applying Holm's and Hochberg corrections.**

Table 8: Random-effects ordered logistic regression models on the willingness to grant permissions in Study 2.

	(1) Base Model	(2) With Controls
Dependent variable: Would you allow or deny the app to access your device's permission X? (Q7)		
Contextualized Justifications experimental group (control group is omitted)	0.077	0.023
	(0.37)	(0.11)
Non-Contextualized Justifications experimental group	-0.064	-0.125
	(-0.28)	(-0.55)
Q8: App functioning w/o accessing permission X	**-0.504*** **	**-0.504*** **
	(-12.21)	(-12.21)
Q9: Privacy dangerousness of permission X	**-0.400*** **	**-0.399*** **
	(-10.86)	(-10.88)
Q10: Negative effects on performance of permission X	0.012	0.008
	(0.27)	(0.17)
Q11: Perceived informativeness of permission X	**0.140*** **	**0.140*** **
	(3.80)	(3.80)
Q3: General app understanding	**0.210** **	0.191**
	(3.01)	(2.76)
Q1: Definition of AR correct		-0.359
		(-1.73)
Q2: AR features on smartphones used		0.353
		(1.73)
Q12_1: Gender - male ("female" is omitted)		0.084
		(0.42)
Q12_2: Gender (prefer to self-identify)		**-1.118** **
		(-3.43)
Q13: Age		-0.005
		(-0.53)
Q15: Education		0.072
		(0.94)
Q16: Technically experienced		-0.170
		(-0.83)
Q17: Smartphone used once or several times a day		**-1.234*** **
		(-3.92)
Q18: Mobile OS (1=Android)		-0.398*
		(-2.21)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; t statistics in parentheses. **Values in bold font indicate statistical significance that hold after applying Holm's and Hochberg corrections.**

Table 9: Ordered logistic regression models on intentions to download the app in Study 2.

	(1) Base Model	(2) With Controls
Dependent variable: How likely are you to download this app? (Q20)		
Contextualized Justifications experimental group (control group is omitted)	0.070	-0.048
	(0.26)	(-0.17)
Non-Contextualized Justifications experimental group	0.134	0.025
	(0.49)	(0.08)
Q8: App functioning w/o accessing permission X		
Index 1	0.038	0.067
	(0.55)	(0.97)
Index 2	-0.113	-0.121
	(-1.36)	(-1.47)
Q9: Privacy dangerousness of permission X		
Index 1	-0.121	-0.122
	(-1.26)	(-1.24)
Index 2	**-0.213****	-0.198*
	(-2.88)	(-2.49)
Q10: Negative effects on performance of permission X		
Index	0.001	-0.053
	(0.01)	(-0.46)
Q11: Perceived informativeness of permission X in understanding app functions and data use		
Index 1	0.081	0.074
	(0.90)	(0.81)
Index 2	0.106	0.098
	(1.34)	(1.18)
Q3: General app understanding	**0.401****	**0.431****
	(3.08)	(3.19)
Q1: Definition of AR correct		-0.360
		(-1.31)
Q2: AR features on smartphones used		0.569*
		(2.18)
Gender - male ("female" is omitted)		0.011
		(0.04)
Gender (prefer to self-identify)		-0.973
		(-1.69)
Q13: Age		0.001
		(0.12)
Q15: Education		0.174*
		(2.00)
Q16: Technically experienced		-0.110
		(-0.42)
Q18: Mobile OS (1=Android)		-0.085
		(-0.37)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; t statistics in parentheses. **Values in bold font indicate statistical significance that hold after applying Holm's and Hochberg corrections.**

D Codebook

Table 10: Codebook for the additional information that participants thought would help them understand what functionalities and data on their device the app will be able to use in Study 1.

Code	Description
No information needed (Q4_2)	Participants do not need additional information, the given information given is clear or sufficient.
N/a	Participants do not offer to add any information (but they do not say the provided information is sufficient like in the "No information needed" code.
General app's functionalities	Functionalities of the apps, not related to privacy/security, e.g. how does the app measure distance, where does the furniture come from, etc.
Instructions	Manual, instructions, help page, FAQ, tutorial
Resources used by the app	Data usage, memory
Clarification about sensors / features (Q4_3)	Definitions of terms, explanations of what the sensors and features are. Indicates insufficient information (when participants provide more details about what kind of information they need to know, e.g. when they require the clarification of how the data collected by these sensors is used, it is "Why/how data is used (purpose)" code).
Possibility to deny/ restrict individual permissions (Q4_4)	Is it possible to deny individual permissions; are the permissions optional/mandatory.
Impact on functionality	How denial of access permissions would affect the functionality (Note: comments about the general app's functionalities are in the "General app's functionalities" code).
Privacy Policies / Terms of services	Privacy policy, terms of services. Includes Privacy Rating / Privacy Ranking.
Privacy concerns (explicit)	Not comfortable with giving access to certain things.
(General) Data handling information	Participants want to know what will happen to the data without specifying whether they are interested in processing, storage, or use conditions.
What data is collected	What specific piece of information are collected.
When data is collected/accessed (Q4_5)	When the data is collected or accessed; common example – while app not in use or all the time.
Where/how data is stored	How data is stored, how long, where it is stored, is it stored at all.
Why/how data is used (Q4_1)	Purpose for data collection, how it is used, how it is processed. Is this data actually required, or is it optional.
How data is shared	Whether the data is going to be shared and with whom (e.g. marketers are mentioned a few times). Specifically, whether the data is going to be sold to the third parties is mentioned often.
How security of my data is ensured	Is the app going to make sure the collected data is secure, and if so how. What security and privacy mechanisms (e.g. anonymization) do they use.
Brief / shorter	When participants mention that the want a short/brief description, or when they want the description to be shorter (note: separate code for "Simpler").
Simpler	Too much detail, or too technical/difficult/jargon-y language. When participants wish the description to be simpler.
Visual	Visual representation, demo, images, video, etc.
One of the specific permissions	Whenever this specific functionality is mentioned (e.g. LiDAR).

PowerCut and Obfuscator: An Exploration of the Design Space for Privacy-Preserving Interventions for Smart Speakers

Varun Chandrasekaran, Suman Banerjee, Bilge Mutlu, Kassem Fawaz
University of Wisconsin-Madison

Abstract

The pervasive use of smart speakers has raised numerous privacy concerns. While work to date provides an understanding of user perceptions of these threats, limited research focuses on how we can mitigate these concerns, either through redesigning the smart speaker or through dedicated privacy-preserving interventions. In this paper, we present the design and prototyping of two privacy-preserving interventions: 'Obfuscator' targeted at disabling recording at the microphones, and 'PowerCut' targeted at disabling power to the smart speaker. We present our findings from a technology probe study involving 24 households that interacted with our prototypes; the primary objective was to gain a better understanding of the design space for technological interventions that might address these concerns. Our data and findings reveal complex trade-offs among utility, privacy, and usability and stresses the importance of multi-functionality, aesthetics, ease-of-use, and form factor. We discuss the implications of our findings for the development of subsequent interventions and the future design of smart speakers.

1 Introduction

Smart speakers, or network-connected speakers with integrated virtual assistants, are becoming increasingly pervasive in households. In 2020, nearly 90 million US adults used a smart speaker [44]. Smart speakers offer their users a convenient way to access information, set alarms, play games, or set to-do lists. Smart speakers also integrate with other devices to realize smart home applications. However, this convenience comes at a potential privacy cost; these devices operate in an always-on mode at earshot of nearby conversations.

Figure 1: The Obfuscator design probe next to a Google Home Mini Device. Obfuscator uses ultrasound jamming to prevent the smart speaker from listening to the user's conversations and is designed to appear as a tabletop "trinket" to blend into the user's home environment.

Smart speakers already provision built-in privacy controls; they are supposed to process audio inputs locally until they detect a wake word, and they pack a button that mutes their internal microphone. Unfortunately, both provisions are not very effective at protecting the user's privacy. Recent incidents raise concerns about *passive* privacy threats [1, 22, 26, 27]. Smart speakers can be mistakenly triggered without the presence of a wake word [11, 19, 25], causing it to record speech not intended as commands. Further, security researchers have documented *active* vulnerabilities that indicate the potential for malicious exploitation of smart speakers [10, 16, 18, 34, 48]. Further, the effectiveness of the mute button to address these problems is in doubt [26]. Recent studies, including the one in this work, indicate that users find this button inconvenient to utilize and not trustworthy in some cases [42]. While different technical interventions have been proposed recently [8, 42], the design space for such interventions remains under-explored. This paper contributes to an improved understanding of the design elements and understanding user

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

experience with these interventions.

In our work, we aim to understand better the user perceptions around the potential technological solutions to the privacy issues involving smart speakers through a technology probe-based approach [21]. The objective of our study is not to validate particular design choices but to understand user perceptions of such interventions better and extract design requirements for them. We utilize the smart speaker's *built-in* mute button as a baseline, to understand user perceptions of how device manufacturers provide privacy control. We utilized two technology probes to represent *bolt-on* privacy-preserving interventions: (a) PowerCut, a smart plug that allows the user to engage/disengage the power supply to a smart speaker remotely, and (b) Obfuscator (Figure 1), which uses ultrasound to deafen the smart speaker's microphone, preventing the smart speaker from listening to nearby conversations. The probes intercept two key resources required for successful smart speaker functionality: *power* (for basic operation) and *microphone inputs* (for voice-based interaction).

To promote user reflection on our privacy-preserving interventions, we conducted in-home demonstrations of our technology probes through in-depth interviews at 24 households. Our interviews took place over two phases between July 2018 and August 2019 , providing us with insight into how such perceptions and attitudes might change over time. Our interviews involved users with diverse demographics, including *casual* (or recreational) users and *power* (or proficient) users, enabling us to distinguish perceptions and design requirements for different user groups. Our findings highlight a complex trade-off between privacy, utility, and usability: the interventions (a) should be plug-and-playable *i.e.,* require minimal setup and upkeep, (b) have a small physical footprint and fit within its environment, (c) offer additional features beyond privacy preservation, (d) does not affect the interaction model with the smart speaker, and (e) must survive the test of time *i.e.,* it should be compatible with existing and future iterations of smart speakers. Through this work, we present the design of our technology probes and our in-home study, and discuss our findings. We conclude with a discussion of their implications for the smart speakers as well as other privacy-sensitive technologies.

2 Background

Our study considers smart speakers deployed in home environments, focusing on (a) Google Home Mini and (b) Amazon Echo Dot as described in Table 1. Users interact with these devices to achieve a multitude of tasks, such as information access, interaction with other smart devices, setting alarms/timers, and voice calls. A typical interaction with a smart speaker starts with the user speaking a wake word, such as "*...Alexa*" or "*...Google.*" Upon recognition of the wake word, the device indicates its readiness to receive command through a visual cue. Then, the device sends the speech segment to the cloud, which verifies the wake word and processes the accompanying command [40]. Verification is necessary since on-device models are typically less accurate to minimize their compute footprint and latency of predictions [32, 41]. As such, the smart speaker has to be always on, continuously listening for a user to speak the wake word. Ideally, the device should only record, and communicate to its cloud, the commands that were triggered by a wake word. In many circumstances, however, the device's operation might not match its expected behavior. This results in the two privacy threats described below. Note, these threats also provide context about scenarios where we envision privacy-preserving interventions to be used.

Feature	Home Mini	Echo Dot
Manufacturer	Google	Amazon
Height × Diameter	4.3 × 9.9 cm	3.3 × 7.6 cm
Wake words	"... Google"	"... Alexa"
Visual Cue	Dots on the surface	LED band
Privacy controls	Mute button	On/Off switch

Table 1: Salient features of smart speakers in 2019.

Passive Threats: The first threat occurs due to innocuous and inadvertent recording *i.e.,* when the smart speaker misunderstands ongoing conversations to contain the wake word. Recent analysis [11] reported that everyday phrases, such as those from TV shows, can accidentally activate a smart speaker, resulting in 10 seconds of speech being sent to the cloud. There have been several incidents where these devices have exported user conversation, including those not preceded with a wake word. While one organization claims this is a one-off act [19], another blames erroneous code [10, 15]. There have also been reported instances where several organizations hired human contractors to listen and tag different recordings from these devices, which include commands and non-commands [16]; this is a severe deviation from perceived device operation. Collectively, we refer to these violations as *passive* privacy threats.

Active Threats: The second occurs due to compromise of the actual device or its operation. A malicious entity can compromise the software running on the connected smart speaker to turn it into a listening device. Such an entity can also change the operation of the device through developing applications that record the user's conversations [25, 34, 49] or inject stealthy commands to wake up the device without the user's awareness [6, 38, 48]. Since these devices are connected to the internet, such alterations are capable of extracting various forms of sensitive information. We refer to such threats as *active* privacy threats.

3 Methodology

We envision privacy-preserving interventions to address potential passive and active threats, especially in scenarios that users perceive as sensitive. Such scenarios can include users receiving visitors or having sensitive conversations. Concretely, there exist two strategies to safeguard users' privacy in sensitive scenarios: (a) redesigning the smart speaker to provide provable privacy guarantees, or (b) designing interventions that co-exist with the smart speaker. The former is a challenging proposition as most of the software and hardware required for successful smart speaker functioning is proprietary. Additionally, it would involve trusting the device provider (a theme that will revisit later) to provide proof that the user's privacy was not violated.

To this end, we explore the design of *bolt-on, hardware-based* interventions. These interventions are less abstract than software-based ones; they allow the users to physically and directly interact with them. For thoroughness, we compare and contrast our findings with the usage of a *built-in* feature found in smart speakers— the mute button. The results of our research inform the design of smart speakers with improved privacy properties and privacy-preserving interventions in physical spaces. Note that our analysis is restricted to smart speakers and not smartphones (which are also susceptible to the threats discussed earlier). In particular, smart speakers are *easier to protect* as they are less *mobile* than smart-phones.

What is a tech probe? We follow a *technology probe*-based design approach, which allows us to identify design guidelines that capture the users' mental models. We aim to understand how the users of smart speakers react to different privacy-enhancing technologies using proof-of-concept prototypes (or probes). In a technology probe, the researcher develops an interface that packages the core functionality of the privacy intervention. The researcher keeps the interface as simple as possible to avoid making design decisions [5, 33]. When an individual interacts with this basic interface, the researcher *probes* the individual to reveal a specific phenomenon that is otherwise hidden [21].

In our case, we probe and interview the users to elicit their immediate reactions and reflections about what design elements are missing and need to be introduced. We follow with qualitative analysis to reveal the design guidelines for a privacy intervention in the smart speaker environment. In follow-up work, we are planning to realize the privacy intervention and set up a diary study to understand longer-term use. This will allow us to concretely measure any issues users have with the actual intervention that was conceptualized for deployment. A note on the nomenclature: in this work, we design technology probes (or probes for short) to elicit insight about the final intervention (which we do not design), for which we make recommendations.

3.1 Iterative Design Process

In designing our technology probes, we followed an iterative design process. We first explored the broad space of solutions (presented in Table 2), their efficacy against an adaptive adversary, and discussed the advantages and disadvantages of each approach. Recall that our objectives are to design an easy-to-use intervention with intuitive yet provable privacy guarantees. It is clear that modifying device hardware and controlling network flow does not provide the desired privacy protection – the encrypted nature of network traffic makes it difficult to tag and discard packets (with information) that are not to be shared, while inadvertent smart speaker activation will persist. One could change the wake word to reduce the frequency of spurious activation/recording. However, this phenomenon is not well understood for it to be a definitive fix, and a harder-to-pronounce wake word has usability problems.

Possible Solutions	***Active* Threats**	***Passive* Threats**
Network interception	✗	✗
Hardware modifications	✗	✗
Change the wake word	✗	✓
Discard smart speaker	✓	✓

Table 2: Space of possible solutions and their effectiveness against malicious programming (or *active* privacy threats) and inadvertent recording (or *passive* privacy threats).

Observe that while some of our possible solutions are intuitive to the average user, others (such as network monitoring) are not. Based on preliminary discussion with several end-users, we converged on a set of dimensions that we found relevant to the final design of our probes. They are (a) the method of user-probe interaction *i.e.,* hands-free vs. physical, (b) the ease of deployment, and (c) the ease of understanding the privacy properties the probe provides. We stress that these dimensions are not exhaustive and merely serve as a starting point for our design.

We construct two probes guided by these suggestions. Again, we stress that we do not seek to evaluate the efficacy of these probes in preserving privacy. We do not attempt to understand how people use these probes as well. Doing so requires running a diary study with the probe deployed in users' homes. We describe the probes used in our study, including those we conceptualized, below. We also briefly state our analysis of the trade-offs ensuing from each probe.

1. Mute: The "mute" feature represents a *built-in* privacy control (Figure 2a). It is available as a push button on the top panel of some of the Amazon Echo Dots and as a sliding button on the side of some of the Google Home Minis. The device manufacturers state that the microphone is deactivated when the mute button is turned on (*c.f.* Figure 2a). Naturally, activating the mute button stops the smart speaker from re-

sponding to the user's voice commands. Upon activation, the Echo Dot's ring color changes to red, and the four lights atop the Google Home Mini turn red.
Trade-offs: While inbuilt, the mute feature requires the user to physically interact with the device to engage the control. It also requires the user to place trust in the manufacturer's implementation of the feature.

2. PowerCut: While the mute button focuses on disengaging the microphone inputs, we conceptualize another probe to disengage the electricity supply. A naive way of achieving our goal is to either disconnect the smart speaker's cord from the outlet or disconnect the cord connected to the smart speaker. However, both options involve physical interaction with the device. Thus, we use a remote-controlled outlet[1] (Figure 2b). The user deploys PowerCut by connecting the smart speaker to the outlet through the smart plug (as seen in Figure 2b).
Trade-offs: We use a commercial smart-plug because we believe that users will be familiar with such products, minimizing their time for acclimatization. Additionally, we speculate that users will trust the functionality of such widely-used products, with no negative publicity. PowerCut is conspicuous and rugged; we believe that its form factor makes it easier to understand and use. The user can engage/disengage PowerCut through a remote control (with a range of operation of 100 feet) without the need to physically interact with the device. Additionally, the smart plug we chose provides a visual cue — an LED glows *red* when powered on to indicate that the smart speaker is active. Clearly, PowerCut offers immediate privacy guarantees. This comes at a cost; the users have to wait for a lengthy boot time whenever they wish to reuse the smart speaker. Additionally, the form factor of PowerCut makes it difficult to use in some environments (with concealed/narrow outlets).

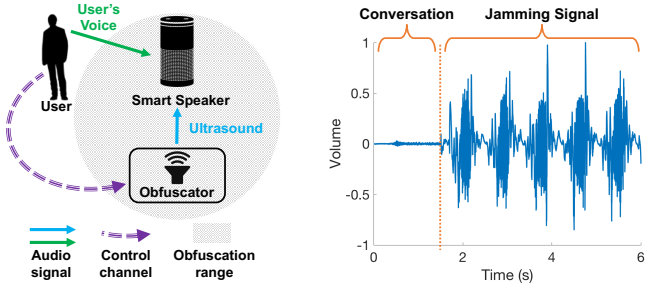
3. Obfuscator: This probe targets the microphone of the smart speaker (Figure 2c). Obfuscator generates *inaudible* ultrasound to deafen the microphone of the smart speaker when the user needs privacy protection (Figure 3a). Using a remote control, users are able to engage/disengage the probe without having to physically interact with it. When disengaging the jamming, the user can *immediately* interact with the smart speaker. Due to non-linearities in off-the-shelf microphones' power and diaphragm [13,37,38,48], Obfuscator creates high-power, human-inaudible noise at these microphones but does not affect its operation. Figure 3b shows the captured signals from a commodity microphone before and after Obfuscator is engaged. Before jamming is invoked, the microphone records a conversation, which is audible at playback. After engaging Obfuscator, the ultrasound jamming signal is recorded at the microphone and completely overwhelms the conversation's signal. The circuitry of Obfuscator includes a remote-controlled DC power supply, an ultrasound generator, and a horn speaker that emits the ultrasound signal.

The design of Obfuscator utilizes a jamming signal with randomized tones at the ultrasound frequency range, which manifest as randomized tones at the audible range. Theoretically, a determined smart speaker manufacturer can attempt to filter these tones at the expense of a degraded speech signal; such degradation might result in a deteriorated performance of wake word detection, which hinders the utility of the smart speaker. Our experiments show that the jamming from Obfuscator is effective at blocking the wake word detection.

[1] Beastron Remote Controlled Outlet

(a) Mute (b) PowerCut (c) Obfuscator

Figure 2: The three employed privacy probes.



(a) High-level overview (b) Microphone received signal

Figure 3: The system design of the Obfuscator probe.

3.1.1 Design Evolution

We explored different design options for the prototype that houses the circuitry. A challenge in prototyping Obfuscator was the footprint of the circuitry. Additionally, horn speakers are bulky, and reducing their size inhibits their efficacy. Our design process started with a search for a privacy metaphor, one that creates the perception of privacy control for the users. Our initial prototype was based on a "cage" metaphor. Here, the Obfuscator probe is housed in a cage-like structure with a door, and the smart speaker is placed within the cage. When the user closes the cage door, Obfuscator generates the ultrasound obfuscation signal to prevent the smart speaker from listening. The user has to manually open the door to disable obfuscation and communicate with the smart speaker. Closing the door "locks" the device in a cage, providing a user with a perception that the device is not active and their space is

private.

The first version of the Obfuscator probe followed the cage metaphor as a 3D printed cylinder (Figure 4a). The cylinder has two compartments; the lower chamber containing the circuit and the ultrasound speaker. The upper chamber has space for the smart speaker as well as the door. The first version has a height of 15.5 cm and a diameter of 12 cm. We refined this design into a lighter and less conspicuous 3D-printed cylinder (Figure 4b) with a height of 13 cm and a diameter of 11 cm. This was the second version.

Based on pilot studies with 2 participants, we found both versions to be neither user-friendly nor fitting with home decor. Participants explicitly indicated that this design was not something they would want in their homes. Further, we observed that individuals did not associate with the privacy metaphor. First, they did not favor the idea of physically interacting with the prototype as it takes away the convenience of using a hands-free device. Second, covering the smart speaker inside the cylinder deprives the users of the ability to observe the visual cue (refer Table 1). This is a shortcoming of placing the smart speaker within the probe. Finally, they thought that the actions of opening and closing the prototype door were conspicuous and would rattle others in the vicinity.

In the third version, we considered three aspects that the users were not fond of: physical interactions with the door, covering the smart speaker, and the aesthetics[2]. The third version of the prototype (Figure 4c) features a platform-like solution, which addresses those shortcomings. This version has a glass cylinder that houses the circuit and is covered by decorative sand; its height is 11 cm, and its diameter is 12.5 cm. The platform, where the smart speaker sits, is encased with synthetic leather. The user can engage/disengage the jamming signal via remote control, obviating the need for physical interaction. This version of the Obfuscator probe follows a different privacy metaphor: "virtual veil." By engaging the jamming signal, Obfuscator creates a virtual privacy dome around the smart speaker, preventing it from listening to the conversations. Our subsequent discussions and reflections about this version revealed that the open nature of the prototype might not enforce the privacy metaphor; users are less likely to perceive privacy control over the smart speaker. Additionally, this version remains co-located with the smart speaker, increasing its form factor. This is not ideal when the smart speaker is concealed.

[2]Aesthetics are subjective, and determining a good aesthetic for even a prototype is a challenging problem.

(a) V.1 (b) V.2 (c) V.3

Figure 4: The design evolution of the Obfuscator probe.

The design search process led to our final prototype of the Obfuscator probe, as shown in Figure 2c. We substantially reduced the form factor of the final version. The new prototype houses the same circuitry in a glass candle holder. The glass is filled with decorative sands and sealed with burgundy burlap. The user only needs to place the prototype next to the smart speaker. This prototype is built using commonly found household artifacts, enabling it to fit in with the existing decor. The final prototype (henceforth our probe) packages the core functionality of the privacy probe: a jamming device that enforces the privacy metaphor. We kept the prototype as simple and basic as possible to avoid making design decisions [5] that influence our findings. In our study, we use the prototype to elicit participants' reflections about what design elements are missing and need to be introduced.

3.2 In-home User Study

We recruited 24 families (including single individuals) within a 15-mile radius of the UW-Madison campus, utilizing the university mailing list, over two phases. Our first phase, in 2018, included 13 interviews, while the second phase (13 months later, in 2019) included 11 interviews. We use a 2-phased approach to obtain results from a wider variety of end-users; we wished to interview both unaware users (in phase 1) and those familiar with media reports of privacy violations induced by smart speaker, at the time of the interviews (in phase 2). We chose to perform shorter and focused interviews as opposed to longer studies (such as diary studies); the tech probe approach allowed us to capture our many goals related to capturing baseline privacy perceptions, introduce the privacy priming, and gain reflection upon interacting with the interventions.

The results reported in the paper are based on interviews with 30 participants ($P_1 - P_{30}$) from these 24 interviews[3]. Our data coding and analysis started immediately and took place simultaneously with data collection, enabling us to monitor the emergence of new codes and themes and determine saturation. We reached saturation by the 18^{th} interview and collected data from 6 more households to assess how perceptions evolve with time. Our approach exhibits several limitations, the most important of which is the sampling of a relatively (a) *ethnically* homogeneous and (b) educated population; the reported results are less likely to generalize to another population of users. We sought to recruit participants with different backgrounds in age, education, and technological proficiency. Our participant pool comprised 15 males and 15 females. The

[3]Some households had more than one participant.

youngest participant was 12 years old, while the oldest was 67 (with a mean age of 37.4 and a standard deviation of 13.9). The occupations of the participants ranged from students to faculty. The wide spectrum in age and profession enables us to gain feedback from a pool with varied technical knowledge and awareness and offers a breadth of experiences and backgrounds that are useful to analyze user interactions with the interventions.

We conducted all interviews at the participants' homes at a time of their convenience. Each interview lasted 90 minutes on average, and the participants were compensated for their time ($40 per study). The study protocol was approved by our Institutional Review Board. Each interview consisted of three stages, which we elaborate on below.

1. Environment Exploration: The interview began with the participants providing a brief tour of their home. Emphasis was placed on the rooms with smart speakers. Then, the interviewer and the participants convened in the room with the frequently used smart speaker so as to simulate a common usage scenario. After obtaining informed consent, the interviewer first asked the participants to interact with their smart speaker to ensure that it was operating as expected. This was followed by questioning participants about their knowledge/understanding of how smart speakers operate. Then, the interviewer asked more detailed questions about the smart speaker's role in the participant's life. The questions focused on frequency, duration, and the purpose of usage. Also, the questions covered the conversations and activities participants perform around their smart speakers. Then, the interviewer inquired about the participant's degree of trust in these devices (in terms of the potential for their conversations to be recorded) and trust in their manufacturers and hypothetical third parties (with whom the recordings might be shared/leaked). The interviewer asked whether individuals have read the news or heard anecdotes about unexpected or undesirable behaviors by the smart speakers. These questions created the appropriate context to discuss privacy-preserving probes; while our follow-up questions are capable of biasing the participants, we believe that they are essential in creating the right environment to discuss the ambiguous space of privacy issues surrounding smart speakers.

2. Interaction with Probes: In a randomly generated order, the interviewer briefly introduced the probes and explained their capabilities to the participants. The participants were given time to familiarize themselves with the probe and set it up (*i.e.,* reorganize their existing layout, if needed, to find a suitable location to utilize the probe). If this was not possible in the room where the interview was occurring, the interviewer and participants discussed why this was the case and moved to a more convenient location with a smart speaker, should one exist. By setting up the probe themselves, we expected the participants to gain greater familiarity with its operation and various other nuances (which we discuss later). The random ordering of probes across participants helped to reduce ordering effects. In settings with families, the interviewer asked different family members to interact with the probe individually (in the presence of other members). After setting up the probe, the interviewer asked the participant to issue voice commands after engaging/disengaging the probe. At each step, the interviewer probed the participant about their level of comfort with the probe and how it impacts the usability of their smart speaker. The participants were encouraged to envision future use-cases for each probe and stress-test the probe's functionality. After the interaction with each probe, the interviewer inquired about the participant's level of trust in the probe. Based on the nature of the response, the interviewer asked several follow-up questions to determine reasons for high/low levels of trust. The interviewer proceeded to discuss perceived privacy control, trust level, convenience, and aesthetics of the probes. On average, users interacted with each probe for approximately 20 minutes[4]. These interview questions were designed to elicit critical reflections – the primary aim of the tech probe study.

3. Concluding Discussions: The interviewer engaged the participants in an open-ended discussion about the probes and their impact on their privacy. The interviewer allowed the participants to hold and observe our probes before answering any other questions related to the study. Finally, the interviewer compensated the participants for their time and effort.

We recorded the interviews, resulting in over 30 hours of recordings, and took photographs of (a) the probes in action and (b) areas where the smart speaker is typically used. We then transcribed, coded, and analyzed the interviews using a Grounded Theory approach [7, 14]. The coding was performed with two coders working independently. Our coders were in moderate agreement, with a Cohen's Kappa (κ) of 0.57 [43]. We started the analysis with an open-coding stage to identify more than 200 informal codes that define critical phenomena in the interview transcripts. Using these informal codes, we extracted recurrent themes within the transcripts and converged on a set of 88 formal codes. We further refined the formal codes into 15 axial codes. We organized the codes into three major themes as summarized in Table 3. We believe that the value of the agreement is acceptable for our study, based on previous research [29]. Following common practices in qualitative coding [3], disagreements were discussed by the coders, followed by code reconciliation, resulting in an updated codebook.

4 Observations

In this section, we discuss the central themes that emerged from our analysis. In summary, we found that: (1) participants were reluctant in sacrificing the convenience associated

[4]From our experience, the users were able to familiarize themselves with the mode of operation and installation of these probes in this timeframe.

Attitudes towards Smart Speaker
Characterizes the user's (a) nature and awareness, (b) technological know-how, and (c) trust in smart speaker manufacturer.
Attitudes towards Probe
Characterizes the user's (a) interaction preference, (b) comfort-levels with regards to usage, (c) long-term technological preferences, (d) trust attitude towards probe, and (e) aesthetics and physical footprint preference.
Utility of Probes
Characterizes the user's preference with respect to probe's (a) multi-functionality, (b) cost, (c) ability to provide fine-grained control, and (d) mode of operation *i.e.,* proactive vs. reactive,

Table 3: Summary of the extracted themes.

with smart speakers; hands-free interaction was most preferred, and physical interaction was seen as being not ideal; (2) participants expected bolt-on interventions with existing household decor and to offer cues informing them of the state of both the probe and the smart speaker; and (3) participants had a preference for multi-functionality and fine-grained control (per-user and per-device). Several of the observations we make have been reported earlier [1, 20, 26, 27]. Our work reaffirms them and shows that the sample used for the rest of the analysis reveals consistent perceptions as previous work[5].

4.1 User Attitudes regarding Smart Speakers

1. Types of Users: Through our study, we identified two types of users: (a) *casual users* who utilize their smart speakers for setting alarms, asking questions, etc., and (b) *power users* who have integrated the smart speakers with other devices in their homes (such as smart lights, house monitoring systems, etc.). We also observed that most participants in our first interview phase were casual users, and a majority of those in the second phase were power users. This phenomenon could be based on the pervasive availability of various smart home devices. We observed that power users (and those in the second phase) were also more familiar with passive privacy violations and with the potential for active violations. We observed that power users were more willing to adapt privacy-preserving interventions as their households were more tightly integrated with the smart speaker. We also observed that a majority of the participants did not change their conversations around the smart speakers, but a small minority reported feeling conscious of having discussions around them. Similar observations were made in recent works studying the privacy perceptions/attitudes of smart speaker users [1, 20, 26, 27].

[5]These findings resulted from our observations in 2018, predating many of the works cited here.

2. Understanding of Smart Speaker Operation: A minority of the participants was unaware of how smart speaker's operate, *i.e.,* they were unaware that their voice commands were processed off-site. Participant P_5, for example, believed that the smart speakers did *"some local learning but also some more... I think at some point people were involved in [the processing]... I think there's an automated learning that occurs to adjust itself to the household, right?"* Abdi *et al.* reported similar observations about users having incomplete mental models of the smart home personal assistants [1].

3. Trust in Device Manufacturers: Our participant pool includes fractions (a) that believed that these organizations could be trusted, (b) that believed that some manufacturers were not in the business of collecting personal information and can be trusted, (c) that trusted the manufacturers, but believed that any information collected could be leaked, and (d) that trusted the manufacturers as long as there is personal utility gained from disclosing said information. A recurrent theme was participants' comfort in being recorded because they believed they were part of a large pool of smart speaker users. Participant P_{10} explains, *"I mean we're not planning any nefarious capers... like we're very boring people and therefore nothing that we're talking about would be of interest to anyone on the other end of [the smart speaker]."* Other studies have also studied user's trust in the device manufacturers and have reached similar conclusions [20, 26].

4.2 User Attitudes Regardingt Probes

1. State of Operation: Participants believed that the current designs of the probes make it too inconvenient to use the smart speaker. They state that using them makes the interaction with a smart speaker a two-phase procedure: first, check the state of the probe (engaged vs. not) and disengage if necessary, and then interact with the smart speaker. Some participants stated that the probes added a *mental burden* in terms of remembering its state. Participant P_9 said: *"when you were to power it off say how do you distinguish that state [when it has no power when using* PowerCut *] from a wake word doing nothing, like I don't know I unmute this right now ... it looks the same."*

2. Ergonomics: Participants were comfortable with the *usability* of the probes. They were easy to set up and use, and the time taken for the probes to activate is acceptable (almost instantaneous in all cases). However, participants expressed dissatisfaction at the longer boot-up times induced by PowerCut. For example, P_8 stated,*"I would find it especially irritating."* Participants suggested that technologies such as Obfuscator that, when disengaged, make the smart speaker*immediately* available were *ideal*. Some participants were concerned about the generalizability of Obfuscator. They believed that the technology is specific to their smart speakers, and would not extend to future smart speakers or smart speakers made by other vendors. Participants were comfortable using a remote

Figure 5: The placement of an Amazon Echo Dot inside an owl-shaped holder in one of the households.

control but felt that their homes have many remotes that could be easily misplaced. When proposing the addition of another remote, P_5 exclaimed, *"they're all over the place, so many remotes! We can't have another remote."* Some participants suggested moving intervention control to a mobile phone app.

3. Trust in Bolt-On Interventions: Finally, participants trust our bolt-on probes more than the built-in mute button. However, participants suggested that trust in a bolt-on intervention would be low if it came from the device manufacturer or any organization that had a similar business model. Participant P_{13} recommended *"a competing company or just a general company that seems like they're like honest"* could develop the interventions. Participants suggested that bolt-on interventions were easier to debug and were easier to understand. However, participants feel that purchasing one bolt-on intervention for every smart speaker would be expensive.

4. Physical Footprint: Participants were concerned with the physical footprint of our probes. While smart speakers were electronic devices, participants often associate them with decorative items (Figure 5) and invest effort in determining where these devices should be placed. A common example of a description about the Obfuscator solution we received was P_2's description: *"a piling on of devices."* Some participants found it difficult to reorganize other items around the smart speaker to facilitate the probe. Additionally, some participants prefer to conceal their outlets, and PowerCut-like interventions would be inconvenient in such scenarios. Participants were uncomfortable with interventions that involve additional wires (as in the case of Obfuscator). Similar observations were made by Pateman *et al.* [35] in the context of the adoption of wearable devices.

4.3 Utility of the Probes

1. Damage to the Environment: Participants were concerned that Obfuscator would cause harm to nearby animals; questions we received upon presenting the Obfuscator were often like P_2's, *"is [this] going to ... make my dog crazy?"* While we did not observe any agitation/discomfort, the participant suggested that their pets could perceive the ultrasound signals and were not bothered. Additionally, participants were concerned about exposing their smart speakers to ultrasound for a prolonged period of time[6].

2. Cost and Multi-Functionality: Cost was repeatedly discussed; participants suggested that the cost of the interventions should not exceed the cost of the smart speaker. Some participants received their smart speakers as gifts. Consequently, they were unable to establish a value for an intervention; P_6 states, *"that's a really interesting question in the sense that I didn't pay for this in the first place. Maybe that's also another reason that I don't have much investment in using this in general."* On the other end of the investment spectrum, we observed that participants who owned multiple smart devices were invested in safeguarding their privacy and were willing to adopt interventions independent of the cost. Participant P_6, who had previously stopped using their smart speakers due to privacy concerns, even stated that they would consider using their device once more given that the interventions were *"cheap... I think would have to rival that remote plug-in cost right because ... it has to be like a cheap utility ... or a cheap accessory like that."* Obfuscator could be used in a proactive way *i.e.,* always-on, or in a reactive way *i.e.,* use when needed. Participants felt that a reactive approach, though tedious, would be easier to understand. Participants also believed that cost could be justified if the intervention provided multiple features. This could be achieved by integrating the design of Obfuscator with other home decors, such as lamps, lights, clocks, radios.

3. Multi-user and Multi-device Environments: The final observation we make is an extension to multi-user and multi-device environments; we observed that in some households, some participants preferred to utilize the intervention more than others. Also, different types of users might exhibit different privacy requirements when interacting with smart speakers. In such scenarios, they desired customized usage profiles based on their requirements *i.e., access control per-user*. Recent research has also indicated the need for access control flexibility in multi-user smart homes [47]. Another observation we make is that some participants preferred to have one intervention (like Obfuscator) being used to preserve privacy against a wide range of smart speaker-like devices. In such scenarios, *access control per-device* was desired. Based on the current design of the Obfuscator prototype, meeting both these requirements is challenging and requires further research. One research direction to make access control per-user more feasible is establishing default privacy options depending on the expected user privacy profiles [2], such as owner vs. visitor.

Concrete Recommendations
1. Aesthetics: The interventions should be offered in different forms, shapes, and colors to fit within people's decors and furniture.
2. Physical Footprint: The footprint of the intervention should be small enough to not force a reorganization of the layout of the owner's house.
3. Multi-Functionality: The intervention is better when providing additional functionality (such as a clock) to reduce its footprint and integrate better with home decor.
4. Ease of Deployment and Understanding: Battery-powered interventions are easier to deploy.
5. Ease of Understanding: A proper understanding of the privacy metaphor improves the adoption of interventions.
6. Trust in Technology: Trustworthy interventions are bolt-on, not network connected, designed by a different trustworthy organization, and pose no additional risk.
7. Mode of Interaction: Using the intervention should not change the interaction with the smart speaker. Hands-free interaction is preferable.
8. Informative Cues: Interventions should offer cues that communicate their state. Visual, auditory, or text cues might be applicable depending on the deployment.
9. Cost: The intervention should cost less than the smart speaker.
10. Fine-grained Privacy Control: The intervention can offer per-user and per-smart speaker privacy controls.
11. Awareness: Awareness of privacy violations increases trust in intervention designers.

Table 4: Summary of the identified design guidelines.

5 Design Implications

Based on the findings from § 4, we make concrete recommendations (based on our findings) on how to design privacy-preserving interventions. The design recommendations are along axes specified in Table 4.

1. Aesthetics: We observed the aesthetics of the privacy interventions to be an important issue for our participants. Participants preferred the interventions to match their individual decorating styles (one example is shown in Figure 5). Many participants suggested that the intervention should come in different forms, shapes, and colors, enabling easier integration within their home decor. As individual tastes vary widely, devising a one-fits-all design is challenging. *One possible approach is to explore different design options for different types of users, including shapes, forms, colors, and material*. This approach has been successful with smart speakers, where participants feel comfortable with the aesthetic of the smart speaker. For example, Amazon has four variants of their Echo featuring combinations of forms and fabric colors.

2. Physical Footprint: Since the smart speakers we considered were small and compact, participants preferred a similar physical footprint for the interventions. Participants expressed concerns regarding the size of both PowerCut and Obfuscator, enquiring if a similar functionality could be achieved with a smaller probe. They believed that using Obfuscator (which needs to be proximate to the smart speaker) requires them to significantly reorganize their existing home decor layout. While the form factor of PowerCut can be reduced trivially, doing so for Obfuscator is challenging; the size and shape of the horn speaker in our current probe were chosen to ensure maximum ultrasound distribution and coverage. Extending such a design to (newer) smart speakers that are larger, or have a different orientation for the microphone inputs, will require rethinking the design and form factor. In summary, interventions that require proximity to the smart speaker need to be designed such that their form factor is comparable to the smart speaker. To achieve such a design, *one recommendation is to design the Obfuscator-style intervention as a stand (upon which the smart speaker can be placed), or as an artifact that can be placed above the smart speaker.* In both designs, the intervention will generate a veil of ultrasound around the entire smart speaker (similar to the horn speaker case that we had designed and evaluated).

3. Multi-Functionality: Closely tied to the aesthetics, participants indicated preference toward an intervention (specifically Obfuscator) that offered features beyond privacy-preservation. They suggested that the Obfuscator intervention could be combined with other household artifacts, such as a lamp, radio, clock, which would further improve adoption. Additionally, *multi-functionality provides an alternative avenue for customizing the probe, making it easier to integrate with existing household decoration*. Such products alleviate the social stigma of being labeled as overly privacy-conscious; such stigma is another reason why the adoption of privacy-preserving interventions is currently low.

4. Ease of Deployment: Participants state that they prefer having a solution that is easy to deploy in their homes; the biggest impediment to any intervention similar to PowerCut is its requirement for an outlet. Many participants preferred to conceal the interventions' wiring, and the nearby outlets can be hard to reach. Attaching PowerCut to wall outlets, even once, requires considerable re-positioning of other devices and their wires. Attaching Obfuscator would require an additional outlet, which is not always readily available. One naïve solution would be to split the outlet among multiple devices. Participants suggested that an Obfuscator design capable of operating on batteries would be more preferred, even if this required periodic replacement.

[6] A detailed study is needed to understand the impact of ultrasound on electronic devices.

Recommendation: Combining the above four observations, we recommend designing interventions in one of two forms: (a) a stand to hold the smart speaker, or (b) a sleeve for the smart speaker (refer Figures 7 and 6). Based on some preliminary analysis, we observe that there is a demand for such artifacts based on our analysis of reviews for such products, and we believe such designs would promote adoption. Since the intervention is not operational in an *always-on* mode, it may be battery-powered — doing away with the requirement for an outlet.

5. Ease of Understanding: All participants were able to easily grasp the metaphor associated with PowerCut, but the technology behind Obfuscator proved complicated for some; some users were unfamiliar with how ultrasound induces a deafening effect. Thus, interventions whose operation is easy to explain may be preferred. This is particularly the case because, while Obfuscator is easy to use once deployed, debugging it may pose problems for users who lack a proper understanding of its operation. We also believe that understanding the detriments (if any) of ultrasound towards humans, animals, and other electronics may put users at ease.

6. Trust in the Technology: Participants were more comfortable with technologies that they believe will survive the "test of time," *i.e.,* be useful for smart speaker models in the future. As discussed earlier, trust also stems from knowing that the interventions do not pose any additional risk. Specifically, we observed that (a) participants wanted to know about any detriments introduced by the interventions, such as potential damage to the smart speaker by frequently disconnecting it from its power source or subjecting it to ultrasound; and (b) our current interventions are not network connected and do not present the same risks as the smart speakers. Finally, participants preferred our bolt-on interventions as opposed to the built-in interventions as they were designed by an organization they trusted (more than the smart speaker manufacturers).

Recommendation: Combining the two points stated above, a concrete design recommendation is to communicate the science behind the operation of the PowerCut-style intervention with a more relatable metaphor or through an interactive demonstration of the intervention's operation. By doing so, we are able to provide more intuition on failure scenarios, which can enable more efficient debugging. This process also assuages any fears related to smart speaker damage or possible harm to nearby entities (such as pets).

7. Mode of Interaction: We observed that participants placed a high value on the *convenience* of using smart speakers, which they are not willing to compromise. Thus, interventions that, when engaged, delay the smart speaker operation (as in the case of PowerCut) are not preferred (even though PowerCut provably preserves privacy, and its mode of operation is very easy to understand). Additionally, any form of physical interaction, be it using remotes or buttons, is far from ideal; some participants expressed preference toward using an app on their smartphones.

Recommendation: We believe that future interventions must be designed so as to have minimal disruption to the convenience of the use of these systems. An ideal design would have a voice interface that allows the user to control it as they control their smart speakers. However, such an always-on and listening privacy-preserving solution can have the same pitfalls as smart speakers, and they must be designed in a manner that does not erode user trust; the mechanism to provide privacy (via a voice-interface) must not become a mechanism for exfiltrating sensitive user conversation (as such a mechanism may require to be network connected). For example, they can lack a network interface to provide the users with hard privacy assurances. Another issue that may arise with voice-activated interventions is erroneous activations; understanding how this can be minimized requires additional research.

8. Informative Cues: As stated earlier, some participants concealed their smart speakers and would prefer concealing their interventions as well. Some participants take this notion to the extreme; they believe that any electronic device that does not provide extensive visual information should be concealed. Thus, visual cues are not ideal in all situations. Additionally, participants suggested that the *red* light on the PowerCut intervention suggested that the intervention was broken, as opposed to indicating the state of the intervention. Interacting with the smart speaker when the intervention is enabled helps users determine the state of the smart speaker (operational vs. not), but such an approach is reactive. Participants indicated a preference for a *proactive* approach.

Recommendation: We propose two recommendations for such settings: (a) the state of the intervention (i.e., engaged vs. disengaged) by communicating to a device that is more optimally placed for being viewed (such as a TV) — this can be done using some form of a closed network connection between the TV and the device via Bluetooth, or (b) the intervention provide auditory cues, where the Obfuscator-style intervention can announce using speech or text that the smart speaker is inactive when users try to activate it.

9. Cost: Another factor that impacts adoption is the cost of the smart speaker. A large fraction of our participants owns smart plugs similar to PowerCut, leading us to believe that such an

intervention is affordable. However, the cost of prototyping Obfuscator was $70, exceeding the cost of smart speakers (priced at approx. $30). This cost includes the price of the commodity parts needed to construct the probe. Participants believe that the cost of the intervention should not exceed the cost of the smart speaker; this is especially true if the intervention can provide privacy protection against a single smart speaker. We believe that if such an intervention would be adopted widely, the production costs could be amortized (and thus have no concrete recommendation to make with regards to minimizing cost). Additionally, understanding the engineering requirements to design an Obfuscator-like intervention that provides privacy against various smart speakers located at different parts of a home requires independent research.

10. Fine-grained Privacy Control: Several households owned more than a single smart speaker, and they had members with different (and potentially conflicting) privacy requirements. Thus, we believe that there is a requirement for (a) fine-grained control *per user*, and (b) fine-grained control *per smart speaker*. For the latter, a naïve solution would be to deploy one intervention per smart speaker, but depending on the cost per intervention, such a solution may not scale. Providing per-user control is a more challenging problem; it requires understanding how disparate the privacy requirements are, how frequently users are utilizing a smart speaker together, and how to mitigate conflicts should they arise.

> **Recommendation:** An ideal design would provide privacy protection for more than one smart speaker. This design could be conceptualized as smaller interventions co-located with the smart speakers but controlled centrally (through some form of closed network).

11. Effect of Awareness: Based on our interview questions, we observed the following trend amidst the participants of our interview phases: participants of our second phase are more concerned about the potential privacy threats from the smart speakers (in comparison to the participants of the first phase, who are also concerned). This concern stems from increased awareness, recent smart speaker mishaps, erroneous code used in them, and immoral practices by device manufacturers. Based on our discussion, we observed that participants believe that these issues are not being seriously audited by the device manufacturers. Discussing various loopholes that can be implemented in the built-in interventions in the status quo (*i.e.,* local wake word processing and the mute button) also increased participants' awareness.

5.1 Consolidated Recommendation

We consolidate the design recommendations based on the aforementioned discussion and provide concrete design guidelines.

Aesthetics, Utility, and Accessibility: Obfuscator-like interventions should be incorporated in accessories that users are already adopting, such as stands and holders (the "owl" shown in Fig 5). Since the completion of our work, we have seen an emergence of a market for such accessories. Additionally, the jamming device should be hidden within a device that is multi-functional, privacy being the secondary functionality. Further, the jamming device should be always-on; the user can access the smart speaker through hands-free interactions, such as gesture-based interaction through wireless sensing. A chime can be played to indicate that the smart speaker is currently active.

Cost & Centralized Control: Obfuscator-like jamming systems rely on directionality to enable their functionality. Thus, it is unclear if there can be *one* of such solutions for *multiple* smart speakers in a home environment. However, many such interventions can be controlled through a centralized interface, such as a single remote control or mobile phone app. Future research is required to better understand the requirements of such a control interface and design it.

Building User Trust: To enhance trust in such interventions, video (or other forms of) presentations/materials can be made to indicate that current smart speakers are purported to exfiltrate home conversation through the use of the public internet. Once this is established, we can educate users of the fact that Obfuscator-like interventions are not connected to any network and consequently can not share sensitive (or any other) information. To further strengthen user belief in bolt-on solutions, end-users can be educated about issues with built-in solutions. Notably, make changes to any built-in solution after deployment requires device manufacturers to regularly and reliably share software updates. However, installing such updates is a challenging proposition to even tech-savvy users [36]. Additionally, as the ecosystem of such smart speaker devices is fast evolving, manufacturers will often not provide support to (a large volume of) smart speakers that were deployed in the past [39]. Additionally, information about software updates (needed for built-in solutions) is not easily accessible, resulting in periods of privacy loss [17].

Accessibility in a Multi-User Environment: Since different users may have different privacy requirements, interventions may be designed to operate with different profiles, such as always-on versus selectively turned on. However, choosing the profile may require (a) explicit user interaction with the intervention, which may be inconvenient, or (b) using auxiliary hardware to identify the users [12].

Enhancing Awareness: Finally, we recommend an on-boarding process that educates the users about the potential privacy threats from accidental/malicious activations, the technology underlying the operation of the intervention, and how to utilize the privacy controls. Such an on-boarding process can take place through voice prompts or an external app; it will increase the user's awareness of the privacy issues as well as improve the user's trust in the probe.

6 Related Work

Privacy Perceptions: The methodology of our study is most similar to Zheng *et al.* [50], and Kaaz *et al.* [23]. They attempt to understand the privacy perceptions of users living in homes with various IoT devices. Similar to our work, surveys are carried out in [4, 9, 28], where the authors try to identify the various challenges associated with setting up and using these devices. Zeng *et al.* [46] and Lau *et al.* [26] study smart speaker users' reasons for adoption through a combination of a detailed diary study and in-home interviews. Similar to this work, we observe that smart speaker users are not privacy-conscious because of the lack of value they associate with their conversational data. Along a similar vein, Abdi *et al.* [1] find that users have incomplete mental models of smart speakers. Similar to our work, they use this understanding to present design recommendations. Some of our findings are coherent with those of Malkin *et al.* [27], *e.g.,* participants are unaware that their conversations are being recorded and stored.

We stress that the primary contribution of our work is *not* in ascertaining the privacy perceptions people have about smart speakers (as done in earlier studies). We wish to understand users' perceptions towards privacy-enhancing technologies and to use this insight to guide the design of both smart speakers and such technologies.

Probe Design: Prior research has investigated system-level solutions to these privacy threats. Feng *et al.* [12] propose continuous authentication as a mechanism to thwart privacy issues related to smart speakers. In our previous work, we propose using ultrasound jamming to address stealthy recording *et al.* [13]. In this work, we wish to validate the usability claims made by the above; consequently, we base our intervention design on the above proposals. The works of McMillan *et al.* [30] and Mhaidli *et al.* [31] provide hands-free alternatives. However, the introduction of a camera to measure gaze introduces privacy concerns. This also requires the user to be in the line of sight of the smart speaker, which reduces its usability. Solutions based on pitch and volume [31] also suffer from similar proximity issues and fail to eliminate privacy violations due to accidental activations.

The Alias project [24] is designed to achieve similar goals to ours. This solution constantly plays noise through a small speaker placed atop the smart speaker and stops the noise upon hearing a custom wake word. Their solution differs from Obfuscator in two ways. First, the Alias intervention does not use ultrasound; the reduced form factor is achieved by not using horn speakers, which are crucial for transmitting ultrasound. Second, the Alias intervention obscures the visual cue provided by the smart speaker; such a design is not preferred. Similarly, work by Chen *et al.* [8] designs a wearable intervention. Wearable solutions offer support in some scenarios, *e.g.,* mobile situations. However, they offer poor support for smart speaker due to lack of proximity to the device. Our experiments with ultrasound-based jamming revealed that the direction of the jamming device and the distance to the smart speaker impact its performance. Additionally, Chen *et al.* do not evaluate the user-related aspects of the intervention, such as user acceptance, aesthetics, and trust.

Design Studies: To safeguard privacy and security in the smart home, Zeng *et al.* [47] prototyped a smart home app and evaluated its effectiveness through a month-long in-home user study with seven households; the users are assumed to be non-adversarial and cooperative. They used their findings to guide future designs for smart home applications. To achieve similar goals as ours, but for smart homes (as opposed to smart speakers), Yao *et al.* [45] adopted a co-design approach and designed solutions with non-expert users. We borrow our study methodology from the work of Odom *et al.* [33]; technology probe studies serve multiple purposes related to designing, prototyping, and field testing the interventions.

7 Conclusions

We presented the design and prototyping of two privacy-preserving interventions: 'Obfuscator' targeted at disabling recording at the microphones, and 'PowerCut' targeted at disabling power to the smart speaker. We presented our findings from a technology probe study involving 24 households that interacted with our prototypes, aimed to gain a better understanding of this design space. Our study revealed several design dimensions for the design of privacy interventions for smart speakers, including multi-functionality, trustworthiness, cues, interaction mode, and ease of deployment.

Acknowledgments

We would like to thank Christopher Little and Thomas Linden who helped with the interviews. We would also like to thank Mariam Fawaz who assisted with the photographs of the interventions, and Yilong Li who assisted with the fabrication of Obfuscator. Finally, we would like to thank the anonymous reviewers and our shepherd for their constructive feedback. Varun, Suman, and Kassem were supported in part through the following US NSF grants: CNS-1838733, CNS-1719336, CNS-1647152, CNS-1629833, CNS-1942014, and CNS-2003129 and an award from the US Department of Commerce with award number 70NANB21H043.

References

[1] Noura Abdi, Kopo M Ramokapane, and Jose M Such. More than smart speakers: Security and privacy perceptions of smart home personal assistants. In *Symposium on Usable Privacy and Security (SOUPS)*, 2019.

[2] Noura Abdi, Xiao Zhan, Kopo M Ramokapane, and Jose Such. Privacy norms for smart home personal assistants. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2021.

[3] Jyoti Belur, Lisa Tompson, Amy Thornton, and Miranda Simon. Interrater reliability in systematic review methodology: exploring variation in coder decision-making. *Sociological methods & research*, 50(2):837–865, 2021.

[4] AJ Brush, Bongshin Lee, Ratul Mahajan, Sharad Agarwal, Stefan Saroiu, and Colin Dixon. Home automation in the wild: challenges and opportunities. In *proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2115–2124. ACM, 2011.

[5] Marion Buchenau and Jane Fulton Suri. Experience prototyping. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*, pages 424–433. ACM, 2000.

[6] Nicholas Carlini, Pratyush Mishra, Tavish Vaidya, Yuankai Zhang, Micah Sherr, Clay Shields, David Wagner, and Wenchao Zhou. Hidden voice commands. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 513–530, Austin, TX, August 2016. USENIX Association.

[7] Kathy Charmaz and Liska Belgrave. Qualitative interviewing and grounded theory analysis. *The SAGE handbook of interview research: The complexity of the craft*, 2:347–365, 2012.

[8] Yuxin Chen, Huiying Li, Shan-Yuan Teng, Steven Nagels, Zhijing Li, Pedro Lopes, Ben Y. Zhao, and Haitao Zheng. Wearable microphone jamming. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 2020.

[9] Eun Kyoung Choe, Sunny Consolvo, Jaeyeon Jung, Beverly Harrison, and Julie A Kientz. Living in a glass house: a survey of private moments in the home. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 41–44. ACM, 2011.

[10] CNN. Google admits its new smart speaker was eavesdropping on users. https://web.archive.org/web/20210226070734/https://money.cnn.com/2017/10/11/technology/google-home-mini-security-flaw/index.html, 2017.

[11] Daniel J Dubois, Roman Kolcun, Anna Maria Mandalari, Muhammad Talha Paracha, David Choffnes, and Hamed Haddadi. When speakers are all ears: Characterizing misactivations of IoT smart speakers. *Proceedings on Privacy Enhancing Technologies*, 2020(4):255–276, 2020.

[12] Huan Feng, Kassem Fawaz, and Kang G Shin. Continuous authentication for voice assistants. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking*, pages 343–355. ACM, 2017.

[13] Chuhan Gao, Varun Chandrasekaran, Kassem Fawaz, and Suman Banerjee. Traversing the quagmire that is privacy in your smart home. IoT S&P '18, page 22–28, New York, NY, USA, 2018. Association for Computing Machinery.

[14] Barney G Glaser and Anselm L Strauss. *Discovery of grounded theory: Strategies for qualitative research*. Routledge, 2017.

[15] Google Home Help. [Fixed issue] Google Home Mini touch controls behaving incorrectly. https://support.google.com/googlehome/answer/7550221?hl=en, 2018.

[16] The Guardian. Apple contractors 'regularly hear confidential details' on Siri recordings. https://web.archive.org/web/20210513003110/https://www.theguardian.com/technology/2019/jul/26/apple-contractors-regularly-hear-confidential-details-on-siri-recordings, 2019.

[17] Kashmir Hill. 'baby monitor hack' could happen to 40,000 other foscam users. https://web.archive.org/web/20210527231505/https://www.forbes.com/sites/kashmirhill/2013/08/27/baby-monitor-hack-could-happen-to-40000-other-foscam-users/?sh=352cfe4558b5, 2013.

[18] Kashmir Hill and Surya Mattu. The house that spied on me. http://web.archive.org/web/20210518035909/https://gizmodo.com/the-house-that-spied-on-me-1822429852, 2018.

[19] Gary Horcher. Woman says her amazon device recorded private conversation, sent it out to random contact. http://web.archive.org/web/20210412111205/https://www.kiro7.com/news/local/woman-says-her-amazon-device-recorded-private-conversation-sent-it-out-to-random-contact/755507974/, 2018.

[20] Yue Huang, Borke Obada-Obieh, and Konstantin Beznosov. Amazon vs. my brother: How users of shared smart speakers perceive and cope with privacy risks. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[21] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B. Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, Nicolas Roussel, and Björn Eiderbäck. Technology probes: Inspiring design for and with families. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 17–24, New York, NY, USA, 2003. ACM.

[22] Haojian Jin, Minyi Liu, Kevan Dodhia, Yuanchun Li, Gaurav Srivastava, Matthew Fredrikson, Yuvraj Agarwal, and Jason I Hong. Why are they collecting my data?: Inferring the purposes of network traffic in mobile apps. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):173, 2018.

[23] Kim J. Kaaz, Alex Hoffer, Mahsa Saeidi, Anita Sarma, and Rakesh B. Bobba. Understanding user perceptions of privacy, and configuration challenges in home automation. In *Visual Languages and Human-Centric Computing (VL/HCC), 2017 IEEE Symposium on*, pages 297–301. IEEE, 2017.

[24] Bjorn Karmann. Project alias. http://bjoernkarmann.dk/project_alias, 2018.

[25] Deepak Kumar, Riccardo Paccagnella, Paul Murley, Eric Hennenfent, Joshua Mason, Adam Bates, and Michael Bailey. Skill squatting attacks on amazon alexa. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 33–47, Baltimore, MD, August 2018. USENIX Association.

[26] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. Alexa, are you listening?: Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 2018.

[27] Nathan Malkin, Joe Deatrick, Allen Tong, Primal Wijesekera, Serge Egelman, and David Wagner. Privacy attitudes of smart speaker users. *Proceedings on Privacy Enhancing Technologies*, 2019(4):250–271, 2019.

[28] Faith McCreary, Alexandra Zafiroglu, and Heather Patterson. The contextual complexity of privacy in smart homes and smart buildings. In *International Conference on HCI in Business, Government and Organizations*, pages 67–78. Springer, 2016.

[29] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

[30] Donald McMillan, Barry Brown, Ikkaku Kawaguchi, Razan Jaber, Jordi Solsona Belenguer, and Hideaki Kuzuoka. Designing with gaze: Tama–a gaze activated smart-speaker. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26, 2019.

[31] Abraham Mhaidli, Manikandan Kandadai Venkatesh, Yixin Zou, and Florian Schaub. Listen only when spoken to: Interpersonal communication cues as smart speaker privacy controls. *Proceedings on Privacy Enhancing Technologies*, 2020(2):251–270, 2020.

[32] Assaf Hurwitz Michaely, Xuedong Zhang, Gabor Simko, Carolina Parada, and Petar Aleksic. Keyword spotting for google assistant using contextual speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 272–278. IEEE, 2017.

[33] William Odom, Richard Banks, David Kirk, Richard Harper, Siân Lindley, and Abigail Sellen. Technology heirlooms?: Considerations for passing down and inheriting digital materials. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 337–346, New York, NY, USA, 2012. ACM.

[34] Danny Palmer. Amazon's Alexa could be tricked into snooping on users, say security researchers. https://web.archive.org/web/20210301140435/https://www.zdnet.com/article/amazons-alexa-could-be-tricked-into-snooping-on-users-say-security-researchers/, 2018.

[35] Matthew Pateman, Daniel Harrison, Paul Marshall, and Marta E Cecchinato. The role of aesthetics and design: wearables in situ. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–6, 2018.

[36] John Pescatore. Securing the "internet of things" survey. https://www.sans.org/reading-room/whitepapers/covert/paper/34785, 2014.

[37] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. Backdoor: Sounds that a microphone can record, but that humans can't hear. *GetMobile: Mobile Computing and Communications*, 21(4):25–29, 2018.

[38] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. Inaudible voice commands: The long-range attack and defense. In *15th USENIX Symposium on Networked Systems Design and Implementation (NSDI 18)*, pages 547–560. USENIX Association, 2018.

[39] Bruce Scheier. The internet of things is wildly insecure — and often unpatchable. http://web.archive.org/web/20210520212158/https://www.wired.com/2014/01/theres-no-good-way-to-patch-the-internet-of-things-and-thats-a-huge-problem/, 2014.

[40] Lea Schönherr, Maximilian Golla, Thorsten Eisenhofer, Jan Wiele, Dorothea Kolossa, and Thorsten Holz. Unacceptable, where is my privacy? exploring accidental triggers of smart speakers. *arXiv preprint arXiv:2008.00508*, 2020.

[41] Siddharth Sigtia, Rob Haynes, Hywel Richards, Erik Marchi, and John Bridle. Efficient voice trigger detection for low resource hardware. In *Interspeech*, pages 2092–2096, 2018.

[42] Ke Sun, Chen Chen, and Xinyu Zhang. "Alexa, stop spying on me!": Speech privacy protection against voice assistants. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, SenSys '20, page 298–311, New York, NY, USA, 2020. Association for Computing Machinery.

[43] Anthony J Viera, Joanne M Garrett, et al. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363, 2005.

[44] voicebot.ai. Nearly 90 million u.s. adults have smart speakers, adoption now exceeds one-third of consumers. https://web.archive.org/web/20210503050256/https://voicebot.ai/2020/04/28/nearly-90-million-u-s-adults-have-smart-speakers-adoption-now-exceeds-one-third-of-consumers/, 2020.

[45] Yaxing Yao, Justin Reed Basdeo, Smirity Kaushik, and Yang Wang. Defending my castle: A co-design study of privacy mechanisms for smart homes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 198. ACM, 2019.

[46] Eric Zeng, Shrirang Mare, and Franziska Roesner. End user security & privacy concerns with smart homes. In *Symposium on Usable Privacy and Security (SOUPS)*, 2017.

[47] Eric Zeng and Franziska Roesner. Understanding and improving security and privacy in multi-user smart homes: A design exploration and in-home user study. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 159–176, Santa Clara, CA, August 2019. USENIX Association.

[48] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 103–117. ACM, 2017.

[49] Nan Zhang, Xianghang Mi, Xuan Feng, XiaoFeng Wang, Yuan Tian, and Feng Qian. Dangerous skills: Understanding and mitigating security risks of voice-controlled third-party functions on virtual personal assistant systems. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 1381–1396, 2019.

[50] Serena Zheng, Noah Apthorpe, Marshini Chetty, and Nick Feamster. User perceptions of smart home IoT privacy. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW), November 2018.

Appendix

A Formal Codes

1. privacy awareness vs education = awareness function of education
2. privacy awareness vs education = not equivalent
3. user technical knowledge = high
4. user technical knowledge = medium
5. user technical knowledge = low
6. user technical knowledge = varies in home
7. user education level = high
8. user education level = medium
9. user education level = low
10. user has concern = yes listening
11. user has concern = yes recording
12. user has concern = yes other
13. user has concern = no
14. user trust large orgs = yes
15. user trust large orgs = case by case
16. user trust large orgs = no
17. user trust third party = yes
18. user trust third party = no
19. user type = power user
20. user type = simple user
21. user accepts listening if = choose over recording
22. user accepts listening if = machine only
23. user accepts recording if = utility
24. user solution choice = discard device
25. user solution choice = unplug device
26. user solution choice = mute
27. user solution choice = remote plug
28. user solution choice = obfuscator
29. user believes in intervention = maybe
30. user believes in intervention = no
31. va state listening = wake word only
32. va state listening = yes
33. va state recording = yes
34. va state recording = non human
35. va state issue attribution = bugs
36. va state issue attribution = unaware
37. va state data use = mundane
38. va state data use = nefarious
39. ecosystem factor = space for solution
40. ecosystem factor = utility of visual cues
41. mute aethestic = acceptable
42. mute aethestic = not acceptable
43. mute haptics = acceptable
44. mute haptics = not acceptable
45. mute form = acceptable
46. mute form = not acceptable
47. mute usability = acceptable
48. mute usability = not acceptable
49. mute concern = privacy protection
50. remote plug aethestic = acceptable
51. remote plug aethestic = not acceptable
52. remote plug form = acceptable
53. remote plug form = not acceptable
54. remote plug haptics = acceptable
55. remote plug haptics = not acceptable
56. remote plug usability = acceptable
57. remote plug usability = not acceptable
58. remote plug concern = boot up time
59. obfuscator aethestic = acceptable
60. obfuscator aethestic = not acceptable
61. obfuscator form = acceptable
62. obfuscator form = not acceptable
63. obfuscator haptics = acceptable

64. obfuscator haptics = not acceptable
65. obfuscator usability = acceptable
66. obfuscator usability = unsure
67. obfuscator usability = not acceptable
68. obfuscator concern = animals
69. obfuscator concern = harm device
70. ideal solution interface = voice
71. ideal solution interface = hands free
72. ideal solution interface = app
73. ideal solution integration = built in
74. ideal solution integration = bolt on
75. ideal solution form = minimal
76. ideal solution form = distributed for devices
77. ideal solution aethestic = multifunctional
78. ideal solution haptics = important
79. ideal solution haptics = not important
80. ideal solution ux = minimal interaction frequency
81. ideal solution ux = no downtime
82. ideal solution ux = no single point control
83. ideal solution ux = single point control
84. ideal solution other = all local
85. ideal solution developer = first party
86. ideal solution developer = third party
87. decision factor = cost
88. decision factor = privacy awareness

B Items in the Commercial Market

We provide screenshots of several cases/sleeves used to encase the Amazon Echo smart speaker. Similar products can be found for the Google smart speaker as well.

Figure 6: A case-like enclosing for Amazon Echo, on Amazon.

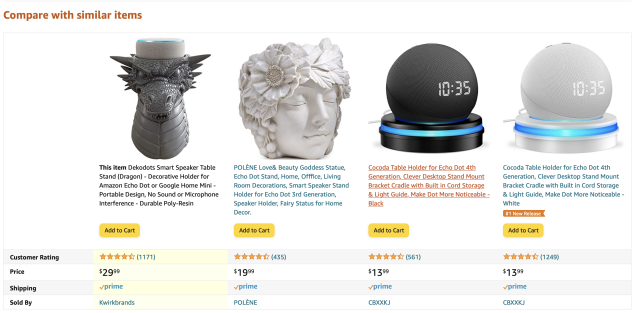


Figure 7: Case-like enclosing recommended by Amazon, for Amazon Echo, on Amazon.

A Qualitative Usability Evaluation of the Clang Static Analyzer and libFuzzer with CS Students and CTF Players

Stephan Plöger
Fraunhofer FKIE

Mischa Meier
University of Bonn

Matthew Smith
University of Bonn, Fraunhofer FKIE

Abstract

Testing software for bugs and vulnerabilities is an essential aspect of secure software development. Two paradigms are particularly prevalent in this domain: static and dynamic software testing. Static analysis has seen widespread adoption across the industry, while dynamic analysis, in particular fuzzing, has recently received much attention in academic circles as well as being used very successfully by large corporations such as Google, where for instance, over 20,000 bugs have been found and fixed in the Chrome project alone. Despite these kinds of success stories, fuzzing has not yet seen the kind of industry adoption static analysis has.

To get first insights, we examine the usability of the static analyzer Clang Static Analyzer and the fuzzer libFuzzer. To this end, we conducted the first qualitative usability evaluation of the Clang Static Analyzer [6] and libFuzzer [16]. We conducted a mixed factorial design study with 32 CS masters students and six competitive Capture the Flag (CTF) players. Our results show that our participants encountered severe usability issues trying to get libFuzzer to run at all.

In contrast to that, most of our participants were able to run the Clang Static Analyzer without significant problems. This shows that, at least in this case, the usability of libFuzzer was worse than of the Clang Static Analyzer. We make suggestions on how libFuzzer could be improved and how both tools compare.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

1 Introduction

The number of critical security vulnerabilities is rising, with the same type of programming mistakes being made over and over again. Testing software for bugs and vulnerabilities is one crucial aspect of helping developers write secure code and countering this development.

The two prevalent approaches for application security testing are static analysis and dynamic analysis.

Static analysis has seen widespread adoption across the industry, dominating the leaders' portfolio of the April 2020 Gartner magic quadrant for application security testing [13]. Dynamic analysis, and in particular fuzzing, has received much attention in academia in recent years, as can be seen by this selection of fuzzing papers published in 2020 alone: [19,29,31–35,38,41–44,46,47,49–51,53,54,58,60–65,67,69,78–87]. Moreover, large software companies such as Google, Microsoft, Cisco and other use fuzzing very successfully, for instance, using fuzzing Google found over 20,000 bugs in Chrome alone [3]. Despite these impressive results, fuzzing has not yet found the same adoption in industry that static analysis has.

In this paper, we examine the usability of the static analyzer Clang Static Analyzer and the fuzzer libFuzzer to get first insights into the question of whether usability issues might be hindering the adoption of fuzzing. For our study, we evaluated several fuzzers and static analyzers. We selected the Clang Static Analyzer because it performed very well in the comparison of Arusoaie et al. [28] and libFuzzer because it is a popular example of a dynamic code analysis tool in academia [46, 55, 66]. However, we would like to stress that neither the Clang Static Analyzer nor libFuzzer are necessarily representative examples of static and dynamic analysis tools. Moreover, since the tools are good at finding bugs of different types, our evaluation should not be seen as a like for like comparison but as gathering first insights into usability strengths and weaknesses of two different tools.

We performed a qualitative mixed factorial design study with 32 CS master students and six competitive Capture the

Flag (CTF) competitors to evaluate the usability of the Clang Static Analyzer and libFuzzer with an easy and a hard task. We designed an easy and a hard task to get a broader view of the tools. The easy task was designed to see if participants could get the tool running in principle while the hard task was designed to reflect a more realistic challenge as would be faced in a real project. The two tools were studied using a within-subjects design to also gather comparative insights of the two tools. The difficulty of the tasks was tested between-subjects with the CS students. The CTF participants only got the hard task. Participants had ten hours over a period of ten days per task to work on the solution.

Our results indicate that the Clang Static Analyzer is easy to use in principle, but it did not scale well to the hard task. Only one CTF participant was able to find the bug, due to a large amount of false positive warnings. With libFuzzer the usability hurdles were much higher, and many CS participants did not manage to solve even the easy task. Even the CTF players did not manage to find the bug in the time allotted although they were able to use libFuzzer in principle. While the majority of participants only failed in the last step of the Clang Static Analyzer, we found usability problems in every step needed to use libFuzzer, which we will discuss throughout the paper.

Supplementary to the Appendix, we provide additional information in a companion document which can be found here: https://uni-bonn.sciebo.de/s/dUH7FOedjHbG5vy.

2 Related Work

The related work section is divided into two parts: usability evaluations of static analysis tools and study methodology concerning developer studies. To the best of our knowledge there are no studies concerning the usability of fuzzers or a usability comparison of static analysis and fuzzing.

Static Analysis Studies Smith et al. [72] conducted a heuristic walkthrough and a user study about the usability of four static analysis tools. They used Find Security Bugs and an anonymized commercial tool for Java, RIPS for PHP and Flawfinder for C. They identified several issues ranging from problems of the inaccuracy of the analysis over workflow integration to features that do not scale. They also conducted a think-aloud study in 2013 with five professional software developers and five students who had contributed to a security-critical Java medical records software system [73, 74]. They wanted to study the needs while assessing security vulnerabilities in software code. The participants worked on four tasks for a maximum of one hour in a lab. However, participants were only asked to examine the reports of the static analysis tool and fix potential bugs but not to run the tool itself. Based on their finding they gave recommendations for the design of static analysis tools. Their main suggestion was that tools should help developers search for relevant web resources. Our study goes beyond this work, since they actually had to use the tool and had more time to do so.

In 2013 Johnson and Song conducted 20 interviews about static code analysis with 20 developers [48]. They found that most participants felt that using static analysis tools is beneficial but that the high number of false positives and the presentation of the bugs were demotivating. In 2016 Christakis and Bird conducted a survey at Microsoft to get more insights into the use of static code analysis [37]. They set the focus on the barriers of using static analysis, the functionality that the developers desire and the non-functional characteristics that a static analyzer should have. They also found that false-positive rates were the main factor leading developers to stop using the analyzer. Developers were willing to guide the analyzer and desired customizability and the option to prioritize warnings. Vassallo et al. confirmed those findings in 2018 [25].

Sadowski and colleges presented a set of guiding principles for their program analysis tool Tricorder, a program analysis platform developed for Google. They included an empirical in-situ evaluation emphasizing that developers do not like false positives and that workflow integration is key [68].

A comparison of open-source static analysis tools for C/C++ code was done by Arusoaie et al. in 2017 [28]. They compared 11 analysis tools on the Toyota ITC test suite [70]. They ranked them by productivity which balances the detection rate with the false-positive rate to compensate for a high false-positive rate. The top three performers were clang, Frama-C [12] and OCLint [20].

Study Methodology Since it is difficult to recruit professional developers [26, 27, 71], Naiakshina conducted a study to evaluate CS students' use in developer studies [57]. They found that for their password storage study students were a viable proxy for freelance developers. Naiakshina followed this up with a comparison to professional developers in German companies [56]. Here they found that the professional developers preformed better overall than students, but that the effects of the independent variables on the dependent ones held none the less and thus conclude that CS students could be used for comparative studies in their case.

A study by Votipka showed that taking part in CTF games tends to have a positive effect on security thinking [76] and hackers are comparable to testers in software vulnerability discovery processes [77].

3 Methodology

We wanted to gain insights into the usability issues of the Clang Static Analyzer and libFuzzer. In the following, we will discuss the design and methodology of the two studies we conducted to do this.

3.1 Tool Selection

We decided to pick one tool per category instead of a spread since it was expected that we would not be able to recruit enough participants to compare multiple tools.

Static Analysis We evaluated the popular commercial static analysis tools Fortify [17], Coverity [9], CodeSonar [7] and checkmarx [4]. Unfortunately, they all forbid publishing evaluations in their terms of use [5, 8, 10, 18]. We based our selection of the open-source static analysis tool on the evaluation of Arusoaie et al. [28]. Based on this, we selected the Clang Static Analyzer, which was the analyzer with the highest productivity rate, a combination of detection rate and false-positive rate, and the highest win rate combining all subcategories within their analysis. We selected the Clang Static Analyzer in version 8.0 as it was the latest version at the time we conducted the first study.

Dynamic Analysis When designing the study, there were no popular commercial fuzzers for C/C++ code available, so we only evaluated the open-source fuzzers: AFL [1], AFL++ [2], libFuzzer, honggfuzz [14] and radamsa [21]. Our literature review showed that AFL/AFL++ and its forks, as well as libFuzzer, are the most common fuzzers in use [30, 31, 36, 40, 46, 52, 55, 78]. Both AFL and libFuzzer were viable choices. While both Fuzzers can fulfil the same tasks, we think that both have strengths and weaknesses for specific situations. To fuzz with libFuzzer a specific function is picked as an entry point. In contrast AFL primarily fuzzes code by using the executable of the target program. In the hard task, it is unrealistic that the code section containing the bug can be reached by AFL this way, while libFuzzer can be run directly on the function. For this reason, we choose libFuzzer over AFL.

3.2 Task Selection

To evaluate the usability of the tools, we needed programs containing vulnerabilities that participants should find. While we were also interested in comparing the usability of the Clang Static Analyzer and libFuzzer, it was not feasible to use the same vulnerabilities for both tools since the types of bugs these tools are good at finding vary too much. We were also interested in comparing how the tools performed at different levels of difficulty. We chose one easy task per tool to get a baseline. With that, we could uncover fundamental difficulties with the tool itself. Additionally, a hard task was chosen per tool to see how it performed in a more realistic setting.

Prerequisites

An appropriate task, i.e. a program to be analyzed, needs to contain a vulnerability that the respective analysis tool can find. This bug should be hard to find by other means than using the tool, particularly by using search engines, thus datasets like the DARPA Cyber Grand Challenge [11] were not viable options for us. We also decided against using tools like Lava-M [39] since they generate a recognizable style of bugs that we knew were familiar to the CTF participants, and inserting bugs into existing programs opens up the risk that participants could use the DIFF tool to identify changes quickly. Ideally, we could use actual undiscovered bugs. To make the matter more complicated, it was also desirable that the difficulty of the two easy and two hard tasks would be similar.

Static Task

We started by running the Clang Static Analyzer on several trending GitHub projects at that time. A list can be found in the Appendix in Table 5. While most of the projects had a high number of warnings, we could not find any true positives, despite investing a significant amount of effort into this. Since this proved fruitless, we contacted experts in static analysis from the Cyber Analysis and Defense and the Cyber Security research departments of the Fraunhofer FKIE to discuss program selection. They did not have any fixed but unpublished bugs, so we were unable to find an unpublished bug suitable for our study. Thus, we fell back on inserting vulnerabilities ourselves but attempting to mitigate the issues mentioned above. For this, we injected one bug in a local copy of the open-source project jq [15] for the easy task and two bugs into a local copy of the open-source project Tesseract [23] for the hard task. The injected bugs were never deployed anywhere outside the study and did not endanger anybody. We chose these projects based on the number of warnings since related work showed that the number of false positives is the main usability issue of static analyzers. Project jq only produced five warnings, and we checked all to confirm they were false positives. Tesseract produced 476 warnings, and we did not find any true positives. We chose to inject one bug, which the Clang Static Analyzer can find without any options activated in both programs. We also injected an additional bug in the hard task, which requires the tester to set the checker *alpha.security.ArrayBoundV2* manually to inspect array boundaries. To mitigate the risk of participants using DIFF to find the inserted bug, we chose older versions of the programs and removed all information concerning the version number. A detailed description of the bugs can be found in the companion document.

Dynamic Task

Unlike with the Clang Static Analyzer, there was no simple way with libFuzzer to evaluate a set of GitHub-projects similarly, so we contacted Code Intelligence a company offering fuzzing as a service to get an overview of difficulty levels of

different projects. Fortunately, they knew a couple of open-source projects with vulnerabilities that had already been reported and fixes submitted but not publicly announced yet. Hence, we selected two of these for our fuzzing tasks. For the easy task, we used yaml-cpp [24] since it is a comparatively small project and has only a handful of public interfaces. This circumstance makes it reasonably easy to get a good overview of the program in a moderate amount of time. Also, writing the fuzz target is relatively simple, and the bug is found in a couple of seconds, even without instrumentation. We knew of one bug in yaml-cpp.

For the hard task, we selected the Suricata [22] project. The fuzz target needed to trigger the bug is more complex than for the yaml-cpp project, and instrumentation, a fitting corpus, and time is needed. Based on the fuzzing expert's recommendations, we opted to give a starting hint to give participants an idea of where they should start looking since the code base was huge, and it would take more time than was available in the study to get an overview. We knew of two bugs in the location where we gave a hint. We fixed one of them since it was a very easy bug, and this was supposed to be the hard task. There were also two other bugs in a different code section. However these were not relevant to our study. So for the purpose of this study, we had one bug in the location for which we gave the hint. In addition to our hint, the Suricata project contained two other sources participants might use to guide their fuzzing effort. The project contained unit tests that could be adapted into fuzz targets. The projects also contained some AFL fuzz targets. As far as we could tell, it was not possible to trigger the bugs with the AFL fuzz targets. Details on the bugs can be found in the companion document.

3.3 Study Design CS Study

Our study contains two independent variables, each with two levels: analyzer (Clang Static vs. libFuzzer) and difficulty (easy vs. hard). Based on our external experts' feedback and internal pre-studies, we decided to allot ten hours for each of the four study conditions. Since this study is highly skill-dependent, we opted for a within-subjects study design for the analyzer variable. To reduce the time needed per participant, we opted to study the difficulty level between-subjects, which then gave us a mixed factorial study design. So each participant either did the easy task with both the Clang Static Analyzer and libFuzzer or did the hard task with both analyzers. We randomly assigned the participants to the hard or easy tasks and randomized the order in which participants used the analyzers to counter learning and fatigue effects. Due to the length of the tasks and the fact that fuzzers need to run for a while to find bugs, we conducted the study online. Participants had ten days per condition and were instructed to work ten hours. If they thought they had found all bugs, they could report in early and would then be given the second task. Participants were asked to keep a diary while working on the task detailing what they spent time on and what problems they encountered. We supplied remote virtual machines with the tools pre-installed for participants to use. They were, however, also allowed to work on their machines if they preferred. After completing both tasks, participants took part in a 30-minute semi-structured interview.

Recruitment and Participants

Since our study required a time commitment of 20 hours, recruiting enough professional developers was not feasible for us at this stage of our research. Thus, we opted to use CS students from a lecture on usable security and privacy and CTF players to gain first insights but want to point out that professional developers would probably perform better in absolute numbers. However, fixing the usability problems discovered by the CS students is likely also to be beneficial to professional developers. However, we cannot make any claims to the extent. Additionally, CS students are also a legitimate user group for these tools, and consequently, fixing usability issues for them is also a desirable goal.

The lecture is part of a master of computer science curriculum and is not mandatory. The focus of the lecture is usability in the context of security. Consequently, all participants had a bachelor degree in computer science, had some basic knowledge on how to evaluate the usability of security tools.

Since the tasks require C/C++ and Linux skills, we used a pre-questionnaire as a filter. We selected a self-reported skill level for Linux and C/C++ of four or higher on a scale of one to seven. We distributed the pre-questionnaire to about 110 students and selected 32 for the study who fulfilled the requirements. They were compensated with an 11% bonus for their end-of-term exam. Students not selected for this study had other opportunities to earn the same bonus. Table 13 of the Appendix shows the demographics of the CS student participants.

Only six out of the 32 participants reported that they have ever used a static code analyzer before. 17 participants, reported that they were familiar with the term fuzzing. However, only four of them had used a fuzzer before. Three of the participants had found a bug with a fuzzer, and one had used the fuzzer libFuzzer before.

3.4 Study Design CTF Study

The second study conducted with CTF players was designed and run after the results of the first study with CS students had been evaluated. While the studies are very similar, we did make three changes which we will highlight here.

Firstly, based on the results of the CS study and the expected skill level of the CTF players, we dropped the easy tasks since we did not expect to learn much there and focused on the hard task.

Secondly, the Suricata project released an update fixing two of the three bugs we knew of, and information about them had been released. To prevent participants from quickly finding these two bugs via a web-search, we gave our participants the updated version, which thus only contained one bug we knew of for them to find. Fortunately, this bug was the one in the code section for which we gave a hint so we could leave the task unchanged except for the update.

Finally, since exam bonus points were not an option, we offered monetary compensation instead. We initially offered a base compensation of 70 euro, with an additional 70 euro offered for finding bugs. We thought due to the competitive nature of CTF players, they would respond well to the incentive. However, we were not able to gain enough interest in our study. After talking to some potential participants, we switched to a flat compensation of 140 euro independent of success.

Recruitment and Participants

We recruited participants from a local Capture-the-Flag (CTF) team via announcements in the weekly team meetings and email. This pool contains roughly 80 people, of which 16 filled out the pre-screening survey. We removed participants who did not have at least one year of CTF experience and had taken part in at least one online and one in-person CTF challenge since we wanted to have a highly skilled group for comparison with the CS students. This left us with eight participants who took part in the main study. During the study, it turned out that two participants had misunderstood the question about in-person CTF events. They actually had not taken part in any and thus are not included in this report.

The demographics of the CTF group are shown in Table 14 in the Appendix. Compared to the CS group, the CTF group is younger, more male and the education level is slightly lower.

Similar to the CS group, only two of the six participants had used a static code analysis tool.

However, all participants reported that they were familiar with the term fuzzing. Five of them had already used a fuzzer before, and three of those five participants had also used libFuzzer and half of all participants reported that they had already found at least one bug with a fuzzer.

This indicates that the CS and CTF group were on a similar level w.r.t. static code analysis tools, but the CTF group had more experience with fuzzing.

3.5 Scoring Results

We evaluated the analyzers based on the success or failure of the participants to get the tool up and running as well as finding the bug. These are separate since it is possible to correctly use the tool but still fail to find the bugs. To make our assessment, we analyzed the submissions of the participants (code and bug reports) as well the content of the diary and the exit interview.

In the static analysis case, a participant successfully fulfilled a static task if the participant used the Clang Static Analyzer correctly and found at least one of the bugs we inserted.[1]

A participant successfully fulfilled a dynamic task if the participant triggered the bug present in the code by using libFuzzer and recognized it as a bug.

4 Limitations

Our studies have the following limitations:

Task Selection. The most considerable limitation concerns the task selection. While we did our best to find fair easy and hard tasks for both tools and consulted external experts, we cannot guarantee that the two easy and hard tasks are exactly the same level of difficulty. While the identified problems likely remain for other tasks, the difference between the two approaches could vary for different tasks.

Participants. We sampled participants from a master course in usable security and a CTF team. Thus this sample is not representative of the wider world. Nonetheless, fixing the issues we found is most likely a good idea, even if more experienced developers might have learnt to overcome them.

Tools Selection. We tested two specific tools: Clang Static Analyzer and libFuzzer. Other tools might perform differently.

Time. Participants only had ten hours per task. While our internal testing suggested that this would be sufficient, some CTF participants would likely have found the bug with libFuzzer, since they were making progress until the end. The time limit did not seem to affect the CS participants or the static tasks.

Unknown Code. Our evaluation only looks at participants analyzing code that they did not write themselves. Further studies with code known to the participants are needed to make claims about this scenario.

Incentives. When comparing the CS and CTF group, the different incentives must be taken into account.

Bugs. During the second study with the CTF participants, information about the bugs in Suricata was published in a blog. One of our participants found this blog and informed us about it. We contacted the author of the blog, and they kindly agreed to take if offline until the end of the study. The participant who informed us about the blog had already finished the task. We asked all other participants whether they had come across the blog or other information online. One additional participant had found some information in an online presentation; however, this did not help them complete the task.

[1] Another true positive bug would also have counted, but this did not occur.

5 Ethics

Our studies were reviewed and approved by the Research Ethics Board of our university.

Our studies also complied with the General Data Protection Regulations. Since we were working with live vulnerabilities, responsible disclosure guidelines were followed. The developers of both programs were already aware of the Bugs, and all participants agreed to comply with responsible disclosure in case they found bugs.

6 CS Study Results

We label participants based on their group (CS or CTF), the order of assignment to the conditions ((FS: fuzzing then static, SF: static then fuzzing) and the difficulty of their tasks(E: easy, H: hard). For the analyses, we used the pre-questionnaire, the reports submitted by the participants, the diaries and the semi-structured post-interview. The questions of the interview and the pre-questionnaire can be found in Appendix A.1 and in the companion document. Except for CS16-FSE, every participant consented to the interview being recorded and transcribed. For the interview of CS16-FSE handwritten notes were taken. The interviews were transcribed and anonymized.

To analyze the interviews and diaries, we used inductive coding [75] with two researchers. The two researchers started with coding the same four randomly chosen interviews independently and in parallel. They compared, analyzed and discussed the two resulting coding sets. It turned out that due to the open approach, the code sets of both researchers were substantially different. Through a discussion of the codes a common coding set was agreed upon. The four interviews were then recoded and discussed again. This procedure was repeated in steps of three interviews. The diaries of the participants were coded with the resulting coding-set from the coding of the interviews. During the coding of the diaries, the coding-set was again supplemented by codes that emerged from the data. All quotes from the participants were translated from German into English by the authors. The final coding set can be found in the companion document.

6.1 Drop-outs

Of the 32 CS student participants, only 18 started the second task, and only ten finished both tasks and were interviewed. CS18-FSH finished both tasks and took part in the interview, but we decided to remove them from our analysis because it became clear that they had not put any real effort into either task. This leaves us with nine participants that finished both tasks and were interviewed. The drop-out rates were much higher than we expected. We have conducted many usability studies with CS students, and it is normal that some drop-out, but this drop-out rate is noteworthy. While we did not conduct formal interviews with the drop-outs, we spoke to some of them. They told us that the tasks were too hard and that they did not know how to solve them and thus dropped out.

The second column of Table 1 shows the drop-out rates, and, as can be seen, only a quarter of the participants dropped out of the easy static task, while half dropped out of the hard static task. With the fuzzing tasks half dropped out both in the easy and hard tasks. This is a first indication that there are usability issues with both approaches. While this explanation seems plausible, based on the rest of the data we could gather, it is also possible that the drop-out rate could be an artefact of our study design. Further studies with different designs are needed to confirm this.

Since we were also interested in a qualitative within-subjects comparison of the Clang Static Analyzer and libFuzzer, most of our analysis focuses on the nine CS participants who completed both tasks and who were interviewed. Table 11 shows an overview of the participants' positive and negative comments and their preference for the two tools. In the following we look at the results in more detail.

6.2 Static Task

The results of the static analysis tasks can be seen in Table 1. Table 6 in the appendix breaks the results down into those who were assigned the conditions as their first or second task. As can be seen, the easy task was indeed relatively easy with only three participants aborting the task. Moreover, in eight out of nine submissions in the easy tasks, the bug was correctly identified. In contrast to that, half the participants dropped out of the hard task. Of those who submitted a report for the hard task, none had found either of the two bugs. In the following, we will group our insights by the different steps needed to complete the task. Readers unfamiliar with this topic can find additional information in the companion document.

Step1: Build Target Program with Clang Static Analyzer
None of the participants reported that they used any other source of information besides the documentation of the Clang Static Analyzer and the target program(TPr).

Not many participants had problems with this step, except for two participants, CS31-SFE and CS24-FSE. Both had used the *configure* and *make* commands on the project to check if everything worked as intended. This interfered with the Clang Static Analyzer because the target program was already built. Therefore the analyzer could not build the target program again and consequently could not find any bugs. CS24-FSE solved this problem on their own. CS31-SFE submitted a report stating that no bugs were found. To gather more information, we let CS31-SFE know that something went wrong and gave a hint. CS31-SFE still counts as a fail in the overall statistics, but with the hint were able to complete this step and their results are considered in the following steps.

Step 2: View the Output Five out of the nine participants who submitted a report had trouble viewing the output of the Clang Static Analyzer. However, this problem only arose because the participants were working on the remote machines offered by us. Except for CS31-SFE, all participants solved the issue by downloading the output to their local machines. Since this problem stemmed from our study setup, we do not see this as a usability issue of the tool.

Step 3: Analyze Reports The presentation of the output of Clang Static Analyzer was rated very positively by the participants. However, as expected, all participants in the hard task and some in the easy task stated that the massive number of warnings was a substantial problem. In particular, the high number of duplicate bug reports was viewed negatively. This is in line with previous work looking at static analyzers. What is noteworthy though is, that this problem has been well known for over a decade but is still an issue with current tools.

6.3 Dynamic Task

The results of the dynamic analysis tasks in Table 1 show that both tasks were hard to solve for our CS participants. For a more detailed overview showing in which order the tasks were assigned, please refer to Table 9 in the Appendix.

Only two CS participants were able to solve the easy task. CS6-FSE dropped out in the following static task, but their diary showed that they straightforwardly solved the task mentioning no problems. The other participant was CS23-SFE, who had stated that they already had experience with fuzzing and libFuzzer in particular. Another participant, CS5-SFE, wrote the correct fuzz target and ran the fuzzer triggering the bug but was convinced that the fuzzing report did not describe a bug.

None of the participants was able to solve the hard task. The drop-out rates for both fuzzing tasks was roughly half, just like for the hard static task.

Unlike with the Clang Static Analyzer, which almost all participants used correctly, we found many problems with the usage of libFuzzer. Table 2 gives an overview of where participants had problems. The columns of Table 2 depict the six steps of the fuzzing process. The first step of finding a suitable function to fuzz contains two values. The first value is the number of functions a participant tried to fuzz. The second value indicates if the participant found a function that triggers one of the bugs known to us. The step of building and instrumenting the target program also contains two values. The first value indicates whether the target program was built, the second if the target program was instrumented. The other columns indicate: how many fuzz targets were created, whether they could build the fuzz targets, whether they ran the fuzzer, triggered the bug, interpreted the output correctly (either as false or true positives), used a corpus and used toy examples to try out fuzzing before trying it on the main project.

The first nine participants in the table are those who completed all tasks and the interview. The next participant in blue is the low effort participant. The participants below in grey completed the fuzzing task but then dropped out. Since we conducted the study online, and participants were allowed to use their own computers, we could not always reconstruct every step. When we were uncertain about whether a participant successfully took a step or not, we marked this with a circle.

For our qualitative analysis, we again focus on the participants that finished both tasks and were interviewed. In the following, we will group our insights by the separate steps needed to complete the task. Readers unfamiliar with these steps can find additional information in the companion document.

Familiarization with the Process All participants started with getting an overview of libFuzzer as well as the target program. Unlike the Clang Static Analyzer, where participants only used the official documentation, many participants searched for additional information about libFuzzer on the web. This highlights deficits in the official documentation as emphasized by CS5-SFE:

> So if you visit the [libFuzzer] page, it is not really obvious what you need to do.

and by CS15-SFE:

> I have not used a fuzzer and I would have wished for a guideline. Such as: Step one, do this, step two, do this... getting started was really hard.

Moreover, participant CS15-SFE stated that the documentation negatively impacted them:

> Even after reading through the paragraphs several times, i'm not sure where to start. Instantly start to losing interest.

Step 1: Find a Suitable Function to Fuzz In the easy task, all participants who identified any functions to fuzz also identified the one that could trigger the bug. Three participants, CS16-FSE, CS31-SFE and CS4-FSH, did not find any functions they thought they could fuzz. CS4-FSH summarized the problems with:

> I looked at the source code of Suricata and was completely overwhelmed. [...] And in the end I did not find any approach how I could fuzz this with a fuzzer.

CS31-SFE commented on that:

> I had problems finding the right point to start fuzzing. The website was not much of a help: [...].

combined	started	drop-out	submitted	success
Static-easy	12	3	9	8
Static-hard	10	5	5	0
Fuzzing-easy	16	8	8	2
Fuzzing-hard	10	6	4	0

Table 1: CS static analysis and fuzzing overview

Participant	Condition	Found Func.	Wrote FT	Build & Inst. TPr	Build FT	Ran Fuzzer	Bug Trig.	Interp. Output	Corpus	Toy
CS16-FSE	easy	✗ / ✗								✗
CS31-FSE	easy	✗ / ✗								✓
CS15-SFE	easy	2 / ✓	2	✗ / ✗ (FT in TPr)	✗					✗
CS24-FSE	easy	1 / ✓	1	✓ / ✗	✗					✓
CS5-SFE	easy	1 / ✓	1	✓ / ✗	✓	✓	✓	✗	✓	✗
CS23-SFE	easy	1 / ✓	1	✓ / ✓	✓	✓	✓	✓	✓	✗
CS4-FSH	hard	✗ / ✗								✓
CS3-SFH	hard	❍ / ❍	❍	✗ / ✗	✗					✓
CS8-FSH	hard	1 / ✗	❍	✗ / ✗						✓
CS18-FSH	hard	✗ / ✗								✗
CS28-FSE	easy	✗ / ✗								
CS6-FSE	easy	2 / ✓	2	✓ / ✗	✓	✓	✓	✓	✗	✗
CS17-SFH	hard	✗ / ✗								✗
CS30-FSH	hard	✗ / ✗		✗ / ❍						
CS26-FSH	hard	❍ / ❍	❍	❍ / ❍						✓

Table 2: CS dynamic analysis deeper statistics: ✓ denotes success in this phase, ✗ failure and ❍ undecidable

> Should I try to look at it from an external view and try to feed information from the outside or should I do it internally [...]? I was missing many examples. It would have been good to not only see somebody fuzzing an easy function [...].

Step 2: Write a Fuzz Target All participants in the easy task who found the function to fuzz also successfully wrote the correct fuzz target. None of the CS participants managed to write a correct fuzz target in the hard task.

Two participants, CS15-SFE and CS3-SFH, tried to write the fuzz target in an existing file of the target program. CS3-SFH changed their mind after having problems with the compilation and used an external fuzz target. For CS15-SFE, this resulted in a more complicated situation. They had to remove the corresponding main function of the target program to use libFuzzer since libFuzzer is shipped with a main function, which interferes with other main functions. More importantly, they also had to modify the make file in order to compile the altered target program. This seemed to have been motivated by the code snipped in the official documentation that could give the impression that the fuzz target is part of the target program. CS3-SFH also stated to this topic:

> I tried to write a simple fuzzer target for a function in app-layer-parser. I started simple and did not manipulate the inputs. I directly wrote it into the app-layer-parser.c file like in the examples given...

CS3-SFH did not include the fuzz target in the report, so we could not confirm this.

While this is a legitimate way to run libFuzzer, in our view writing the fuzz target in a separate file is a cleaner and more straightforward approach.

Step 3: Compile and Instrument Target Program In the case of the easy task, none of the CS participants, except CS23-SFE, seemed to be aware that instrumentation exists, or had any idea why instrumentation is useful. CS23-SFE was the only participant who actively dealt with instrumentation and was aware of the implications of the "fuzzer-no-link"-flag and was the only ever to use it.

All participants of the hard task had the problem that Suricata builds as an executable, and libFuzzer can not directly fuzz executables. None of them was able to find a solution for this, as depicted in the companion document. CS8-FSH tried to find a solution by exporting a function from the Suricata elf binary into a shared object and then load and run it within a fuzz target. However, they were not able to do so.

Step 4: Build Fuzz Target The five remaining participants reported severe problems in the building and linking step. CS24-FSE stated:

> I believe that the library itself wasn't the problem, but the stupefying linking and compiling was.

The problems with building and linking could have a variety of reasons. Two participants stated that they lacked knowledge concerning the make system (CS24-FSE, CS15-SFE) or even compiling C/C++ code in general (CS5-SFE). Other participants had problems linking libraries and were randomly trying out compiler and linker flags to get the fuzz target to compile. For yaml-cpp, some participants also tried to use *make install* on the target program to increase the chance of hitting the right combination of compiler and linker flags. Overall, we observed a lack of understanding concerning the interaction between fuzz target, target program and compilation process.

Step 5: Run and Observe the Fuzzer In the easy task CS5-SFE, CS6-FSE and CS23-SFE were able to build the fuzz target and run libFuzzer. Moreover, they all triggered the bug because in the easy task the bug was triggered within seconds.

Step 6: Interpret Output Of the three participants who triggered the bug, CS5-SFE incorrectly classified the output as a false positive. CS5-SFE saw the out of memory error and the malformed input the fuzzer had generated but thought this was a mistake by libFuzzer instead of a bug in the program.

Even though CS23-SFE was by far the best participant solving the easy fuzzing task in less than two hours, they did not find the output of libFuzzer very helpful, stating:

> I would be helpful if the output did not just contain the input which led to the bug, but also information about the crash.

Toy Examples and Documentation Six of the nine participants experimented with the toy examples from the documentation to get to know libFuzzer. However, as described above, this led some astray.

7 CTF Study Results

The interviews and diaries of the CTF group were coded based on the same principles we used for the CS group. The questions of the interview and the pre-questionnaire can be found in the Appendix A.2 and in the companion document.

An overview of the CTF-group's success can be seen in Table 3. Unlike in the CS group, we had no drop-outs in the CTF group. There are two potential explanations for this. Based on our interviews, the CTF participants were not as frustrated with the tools as the CS participants or had a higher frustration threshold and a willingness to work with complicated and puzzling systems. However, it could also be that the 140 euro incentive was more motivating than the 11% exam bonus or a combination of these factors. As in the previous section, we will structure our results around the steps needed to operate tools.

Static Analysis

Steps 1 & 2 The participants had no problems getting to the point where they had to inspect the reports given by the Clang Static Analyzer. Some participants reported issues viewing the results, like in the CS study, but could quickly solve them.

Step 3: Analyze Reports Overall, participants were satisfied with the usability of the tool as with the presentation of the output but had the same problem with the high number of false positives as the CS group. CTF7-SF stated:

> More than once I wondered whether it's me or the analyzer who doesn't understand the code.

Only one participant (CTF2-FS) was able to find one of the bugs.

Notably, four out of six participants reported that they heavily prioritized reports in the category *memory errors*. Some specifically mentioned that they neglected reports in other categories, such as *Logic errors*, which was the category where the Bug was. Their reasoning was that these kinds of bugs potentially have low exploitability. In the interviews, some of the CTF participants stated that they did not consider availability/denial-of-service an issue in this context. This could be an artefact of the fact that in CTF games denial-of-service attacks are often forbidden. CTF7-SF stated:

> Going through the "Memory error" bugs - If there are any vulnerabilities I expected to find them here, so I took some time for them.

All in all, participants showed strong tendencies to focus on bug types, ignoring much of the output produced by the Clang Static Analyzer. CTF3-SF summarized it as follows:

> I filtered for use-after-free and double free/delete, which seemed most likely to have immediate security impacts. While there were 72 bugs shown in total, most of them were duplicates. I decided to only look at one bug per bug group/function-combination, which eliminates mostly very similar code paths... For each combination, I chose the shortest path length to have a minimum-complexity example of a triggering code path.

This filtering caused the participants to miss our bug, which was in the category *Dereference of undefined pointer value*.

combined	started	dropout	submitted	success
Static-hard	6	0	6	1
Fuzzing-hard	6	0	6	0

Table 3: CTF: overall results

Dynamic Analysis

Despite being more experienced and security savvy, our CTF participants also had trouble with libFuzzer. Table 4 gives an overview of where participants had problems.

Step 1: Find a Suitable Function to Fuzz Unlike the CS participants, all CTF participants were able to identify the correct function to fuzz.

Step 2: Write Fuzz Target The writing of the fuzz target split the CTF group in two. Participants CTF4-FS and CTF5-SF used the unit tests as the basis for their fuzz targets. Participants CTF2-FS, CTF3-SF, CTF6-FS and CTF7-SF based their fuzz target on the AFL targets contained in the project. In general, all participants agreed that creating the fuzz target was a complicated and time-consuming task.

Step 3 & 4: Compilation and Instrumentation Five out of six participants were successful in compiling and instrumenting all necessary parts, only CTF5-SF did not successfully manage this step. CTF5-SF had criticism for the documentation and some suggestions on how the usability of these steps could be improved.

> Although everything was described [in the example] instructions were missing how to approach fuzzing a real-world project, how to integrate it into an existing boot-system. Maybe one could have made something generic to integrate it into Cmake or Auto-build.

Four of the five participants who created a fuzz target wrote the fuzz target directly into the target program. Unlike the CS participants, they were able to make the necessary modification to make this work. We found this interesting since it does not seem to be the intuitive way for us.

Step 5: Running and Observing the Fuzzer The four remaining participants, CTF2-FS, CTF3-SF, CTF6-FS and CTF7-SF, created multiple fuzz targets and observed the fuzzing process.

All participants focused on using the executions per second as well as the code coverage as the indicators on whether the fuzzing process was going well or not. Concerning the code coverage, some participants mentioned that it could sometimes be hard to interpret the relative magnitude of the given value correctly. CTF6-FS summarized it as follows:

> Of course this depends on the complexity [of the TPr], but when I have such a HTTP fuzzer, and I know it is implemented in C, and I only have twenty branches or so which have been covered, then I know: This can't be. This absolutely can't be! You can't implement a HTTP fuzzer with so few branches or so few basic blocks. And if it also isn't making progress, then, you need to find out what is the matter.

The problem of knowing whether libFuzzer covered the necessary parts of the code was a frequently reoccurring statement. Only CTF2-FS used the visualizer of LLVM to get a better understanding of the situation.

Step 6: Interpret the Output CTF4-FS wrote a fuzz target and was also able to build it. However, the fuzz target quality was relatively low, so that the fuzz target crashed directly due to problems during initialization when executed. The participant was aware of the problem but could not fix it. CTF4-FS stated:

> And when I wanted to fuzz the correct filter, I always failed because something was uninitialized and this was why it always crashed. So it always fuzzed but crashed in each attempt.

Unsurprisingly, CTF4-FS believed that fuzz target creation was a big problem. They reasoned that this might partially be because they did not know the code.

> It was probably because I didn't know the software at all and then I couldn't proceed as well as I hoped

All of the four remaining participants were able to interpret the output of libFuzzer. Depending on the situation, they handled the corresponding situation differently.

CTF2-FS and CTF3-SF had problems with memory leaks due to how they implemented the fuzz targets. They were able to fix the problems and re-ran the fuzz target without the memory leak.

CTF7-SF wrote at least ten fuzz targets and ran them. Their fuzz target for the smb protocol crashed for every input. They decided that the fuzz target was flawed and just ignored it

Participant	Found Func.	Wrote FT	Build & Inst. TPr	Build FT	Ran Fuzzer	Bug Trig.	Interp. Output	Corpus	Toy
CTF5-SF	1 / ✓	1	✗ / ✗	✗					✗
CTF4-FS	1 / ✓	1	✓ / ✓ (FT in TPr)	❍	❍	✗	✓		✗
CTF2-FS	1 / ✓	1	✓ / ✓ (FT in TPr)	✓	✓	✗	✓	✓	✗
CTF6-FS	1 / ✓	10	✓ / ✓ (FT in TPr)	✓	✓	✗	❍	✗	✗
CTF7-SF	1 / ✓	11	✓ / ✓	✓	✓	✗	✓	✓	✗
CTF3-SF	1 / ✓	3+	✓ / ✓ (FT in TPr)	✓	✓	✗	✓	✓	✗

Table 4: CTF dynamic analysis deeper statistics: ✓ denotes success in this phase, ✗ failure and ❍ undecidable

because they had several other fuzz targets that were up and running.

CTF7-SF's fuzz target for the dnp3 protocol also produced many errors, but again they understood that this was due to a flawed fuzz target and not because of actual bugs. They attributed the flaws initialization problems and did not fix them for the same reason as before. CTF2-FS and CTF3-SF also had problems with initialization, but both fixed the issues to make the fuzz target work.

None of the four participants found a bug. However, all four were using libFuzzer correctly, and with more time available, it seems likely that they would have found the bug in the target program. While our pre-testing suggested that ten hours was enough time, future iterations of this kind of study should plan more time for this kind of task. Nonetheless, we are confident that they would be capable of finding these kinds of bugs with libFuzzer in the wild with the skill they already possess. However, the effort and skill required are quite substantial. In contrast, we do not believe that our CS would be able to use libFuzzer without investing significant effort in learning how to use the tool.

Expanding the Search As the participants did not encounter any true crashes, they felt the need of exploring further options. Most of them did this by manually targeting specific parts of the code. Still not encountering any crashes, they tried to optimize the fuzz targets and tried to develop more complex inputs to the functions. In the interviews participants CTF6-FS and CTF3-SF phrased this as a feature request.

Consequently, stateful fuzzing was needed. CTF3-SF considered to implement stateful fuzzing but was not able to do it in the given time. CTF6-FS implemented a minimal form of stateful fuzzing. However, they were not very enthusiastic about it:

> Libfuzzer does not support stateful fuzzing, therefore no high expectations as path stability will be horrible.

Corpus and Dictionary Except for CTF2-FS, all participants used corpora for their respective fuzz targets. Interestingly, CTF6-FS used both a corpus and a dictionary. CTF6-FS observed their fuzz targets with a corpus and a dictionary included and noticed a drop in performance because the coverage was lower than without the corpus and dictionary. Consequently, they proceeded without either.

8 Discussion

8.1 Clang Static Analyzer

The Clang Static Analyzer enabled even inexperienced users to check the target project for potential security issues. With the Clang Static Analyzer, both our participant groups were able to start the process reasonably easily and quickly. The usability of the tool was consequently viewed fairly positively. Our participants intuitively used Nielsen's view on usability [59], which separates usability and utility. In the hard task, the high number of false-positive warnings was seen negatively by both the CS and CTF groups, but this did not affect their perception of "usability". The CTF participants also had a negative view of the usefulness in general. They did not think the tool was helpful when looking for vulnerabilities. Consequently, they saw the tools as having good usability but bad utility. It is worth noting though, that under the ISO 9241 [45] definition of usability, the bad effectiveness and efficiency measured against the capability of finding true bugs would lead the Clang Static Analyzer to receive a bad usability evaluation.

Thus, the holy grail of static analysis continues to be the reduction of the number of false positives.This would improve the utility under Nielsen or usability under ISO 9241 and enable users to effectively and efficiently find bugs.

8.2 libFuzzer

In stark contrast to the Clang Static Analyzer, where participants only struggled in the very last step, we found no step in the libFuzzer process that did not cause our participants severe problems. Our CS participants struggled even with the easy fuzzing task showing that the usability of libFuzzer is not at a comparable level to the Clang Static Analyzer. Even our skilled CTF players found many aspects vexing, unnecessarily complicated and burdensome. However, in theory, the utility of libFuzzer is good. Consequently, we see a lot of potential if the usability can be improved.

Based on our observations, our recommendations for libFuzzer are:

- **Assist users in finding suitable functions to fuzz** It would be useful if libFuzzer assisted users in identifying functions worth fuzzing quickly. This was not an issue for our CTF participants, but if libFuzzer is to see the same level of adoption as static analysis, it needs to be usable by non-experts as well.
- **Fuzz-target creation** This is one of the most important points. It takes a lot of expertise to write anything but the most trivial fuzz targets for libFuzzer. In the case of Suricata, participants actually wrote multiple fuzz targets for the same function to account for the different parsers. Either assisting in creating fuzz targets or making the coverage guided self-exploration of libFuzzer more intelligent would be a great benefit. It is essential for less experienced users, but it would also save time and effort for users like our CTF players.
- **Build automation** The building and linking process currently also requires a lot of manual work for non-toy projects, and it also requires a good understanding of how the different components interweave. It would be highly desirable to automate a lot of this, so users do not need to understand, or know of, these issues.
- **Opt-out sanitizers:** Currently the use of sanitizers is opt-in, i.e., the user has to integrate them actively. We would recommend including many of these by default and letting users opt out if necessary.
- **Support automatic stateful fuzzing** Many situations require stateful fuzzing to achieve good performance. In libFuzzer, this is a completely manual task, and some of our CTF participants even wrote their own stateful fuzzers to deal with the situation.
- **Improve Code Coverage** Our study shows that Code coverage plays a major role in the usability of libFuzzer. Even our CTF participants struggled to write fuzz targets that covered all the code of just one target function. This had to be done manually because libFuzzer is not yet powerful enough to do this on its own in a reasonable time. Potentially focusing on code coverage close to fuzz targets would be a worthwhile endeavor to increase usability.
- **Better documentation** Finally, while this is not particularly glamorous and is a well-known problem in many areas, we saw a clear need for better documentation. There is a clear difference between the Clang Static Analyzer and the libFuzzer documentation despite both belonging to the LLVM project. The current libFuzzer documentation led some of our participants astray. In particular, we recommend creating more complex examples instead of just using toy examples.

8.3 Comparison

Since we conducted a within-subjects study, we were also interested in our participants' comparative view of the two tools. To support our impressions from the interviews and diaries, we also analyzed the number of positive and negative comments to get an overview of the disposition towards the two tools.[2]

The majority of CS participants favored the Clang Static Analyzer when answering the question of which tool they would want to use in the future, including those faced with over 500 warnings in the hard task. In contrast, the CTF participants had a somewhat ambivalent relationship to the Clang Static Analyzer. In principle, they described the usability positively and had fewer negative comments for the Clang Static Analyzer than for the libFuzzer. However, they did not see the Clang Static Analyzer as a serious contender to find vulnerabilities. As a result, they stated that they favored libFuzzer for future use and often stated that they would only use the Clang Static Analyzer for fixing style issues.

That is because they saw far more potential for libFuzzer than for the Clang Static Analyzer and thus would use libFuzzer. The corresponding Table 11 can be found in the Appendix.

So, in summary, our interpretation of the results suggests that poor usability of libFuzzer and the good usability of the Clang Static Analyzer led CS students to prefer it despite the poor utility. However, the CTF participants acknowledged the better usability of the Clang Static Analyzer but saw too little utility to want to use it for their work in the future and tolerating the poor usability of libFuzzer due to its better perceived utility.

9 Conclusion and Future Work

In this paper, we presented the first qualitative studies examining the usability of libFuzzer and the Clang Static Analyzer. In the context of our study design, we found that the Clang Static Analyzer offers good usability but poor utility, while libFuzzer offers poor usability but better utility. Since static analysis and fuzzing find different kinds of bugs, ideally, they would both be used in tandem. For this, the usability of libFuzzer would need to be improved to lower the bar for entry. To aid in this, we identified several usability issues in libFuzzer and make suggestions for improvements.

Acknowledgments

The authors would like to thank Sirko Höer for helping with the fuzzing task selection. This work was partially funded by the ERC Grant 678341: Frontiers of Usable Security.

[2]The comment count does not necessarily reflect the weight of individual issues but offers interesting insights nonetheless.

References

[1] Afl. https://github.com/google/AFL. Accessed: 02-13-21.

[2] Afl++. https://github.com/AFLplusplus/AFLplusplus. Accessed: 02-13-21.

[3] Bugs found in chrome with fuzzing. https://bugs.chromium.org/p/chromium/issues/list?can=1&q=label%3AClusterFuzz+-status%3AWontFix%2CDuplicate&colspec=ID+Pri+M+Stars+ReleaseBlock+Component+Status+Owner+Summary+OS+Modified&x=m&y=releaseblock&cells=ids. Accessed: 02-13-21.

[4] Checkmarx sast. https://www.checkmarx.com/de/products/static-application-security-testing. Accessed: 02-13-21.

[5] Checkmarx sast license agreement. https://checkmarx.atlassian.net/wiki/spaces/CCD/pages/1253442222/CxIAST+End+User+License+Agreement+EULA. Accessed: 02-13-21.

[6] Clang static analyzer. https://clang-analyzer.llvm.org/. Accessed: 02-13-21.

[7] Codesonar sast. https://www.grammatech.com/codesonar-cc. Accessed: 02-13-21.

[8] Codesonar sast license agreement. https://support.grammatech.com/documentation/licenses/GrammaTech_License_Agreement_CodeSonar_ver.2016.1.0.pdf. Accessed: 02-13-21.

[9] Coverity scan. https://scan.coverity.com/. Accessed: 02-13-21.

[10] Coverity scan license agreement. https://www.synopsys.com/company/legal/software-integrity/coverity-product-license-agreement.html. Accessed: 02-13-21.

[11] Darpa cyber grand challenge. https://www.darpa.mil/program/cyber-grand-challenge. Accessed: 02-13-21.

[12] Frama-c. https://frama-c.com/. Accessed: 02-13-21.

[13] Gartner magic quadrant for application security testing. https://www.gartner.com/en/documents/3984345. Accessed: 02-23-21.

[14] Honggfuzz. https://github.com/google/honggfuzz. Accessed: 02-13-21.

[15] Jq. https://github.com/stedolan/jq. Accessed: 02-13-21.

[16] libfuzzer. https://llvm.org/docs/LibFuzzer.html. Accessed: 02-13-21.

[17] Microfocus-fortify. https://www.microfocus.com/de-de/products/static-code-analysis-sast/overview. Accessed: 02-13-21.

[18] Microfocus-fortify license agreement. https://www.microfocus.com/media/documentation/micro_focus_end_user_license_agreement.pdf. Accessed: 02-13-21.

[19] Mofuzz: A fuzzer suite for testing model-driven software engineering tools.

[20] Oclint. http://oclint.org/. Accessed: 02-13-21.

[21] Radamsa. https://gitlab.com/akihe/radamsa. Accessed: 02-13-21.

[22] Suricata. https://suricata-ids.org/. Accessed: 02-13-21.

[23] Tesseract ocr. https://github.com/tesseract-ocr/tesseract. Accessed: 02-13-21.

[24] yaml-cpp. https://github.com/jbeder/yaml-cpp. Accessed: 02-13-21.

[25] Context is king: The developer perspective on the usage of static analysis tools. *25th IEEE International Conference on Software Analysis, Evolution and Reengineering, SANER 2018 - Proceedings*, 2018-March:38–49, 2018.

[26] Y. Acar, M. Backes, S. Fahl, D. Kim, M. L. Mazurek, and C. Stransky. You get where you're looking for: The impact of information sources on code security. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 289–305, 2016.

[27] Y. Acar, S. Fahl, and M. L. Mazurek. You are not your developer, either: A research agenda for usable security and privacy research beyond end users. In *2016 IEEE Cybersecurity Development (SecDev)*, pages 3–8, 2016.

[28] Andrei Arusoaie, Stefan Ciobaca, Vlad Craciun, Dragos Gavrilut, and Dorel Lucanu. A comparison of open-source static analysis tools for vulnerability detection in C/C++ Code. *Proceedings - 2017 19th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing, SYNASC 2017*, pages 161–168, 2018.

[29] Cornelius Aschermann, Sergej Schumilo, Ali Abbasi, and Thorsten Holz. Ijon: Exploring deep state spaces via fuzzing. In *2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020*, pages 1597–1612. IEEE, 2020.

[30] Domagoj Babić, Stefan Bucur, Yaohui Chen, Franjo Ivančić, Tim King, Markus Kusano, Caroline Lemieux, László Szekeres, and Wei Wang. Fudge: Fuzz driver generation at scale. ESEC/FSE 2019, page 975–985, New York, NY, USA, 2019. Association for Computing Machinery.

[31] William Blair, Andrea Mambretti, Sajjad Arshad, Michael Weissbacher, William Robertson, Engin Kirda, and Manuel Egele. Hotfuzz: Discovering algorithmic denial-of-service vulnerabilities through guided micro-fuzzing. *Proceedings 2020 Network and Distributed System Security Symposium*, 2020.

[32] Tegan Brennan, Seemanta Saha, and Tevfik Bultan. Jvm fuzzing for jit-induced side-channel detection. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, ICSE '20, page 1011–1023, New York, NY, USA, 2020. Association for Computing Machinery.

[33] Alexandra Bugariu and Peter Müller. Automatically testing string solvers. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, ICSE '20, page 1459–1470, New York, NY, USA, 2020. Association for Computing Machinery.

[34] Hongxu Chen, Shengjian Guo, Yinxing Xue, Yulei Sui, Cen Zhang, Yuekang Li, Haijun Wang, and Yang Liu. MUZZ: Thread-aware grey-box fuzzing for effective bug hunting in multithreaded programs. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 2325–2342. USENIX Association, August 2020.

[35] Yaohui Chen, Peng Li, Jun Xu, Shengjian Guo, Rundong Zhou, Yulong Zhang, Tao Wei, and Long Lu. SAVIOR: towards bug-driven hybrid testing. In *2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020*, pages 1580–1596. IEEE, 2020.

[36] Yuanliang Chen, Yu Jiang, Fuchen Ma, Jie Liang, Mingzhe Wang, Chijin Zhou, Xun Jiao, and Zhuo Su. Enfuzz: Ensemble fuzzing with seed synchronization among diverse fuzzers. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 1967–1983, Santa Clara, CA, August 2019. USENIX Association.

[37] Maria Christakis and Christian Bird. What developers want and need from program analysis: an empirical study. In David Lo, Sven Apel, and Sarfraz Khurshid, editors, *Proceedings of the 31st IEEE/ACM International Conference on Automated Software Engineering, ASE 2016, Singapore, September 3-7, 2016*, pages 332–343. ACM, 2016.

[38] Sushant Dinesh, Nathan Burow, Dongyan Xu, and Mathias Payer. Retrowrite: Statically instrumenting COTS binaries for fuzzing and sanitization. In *2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020*, pages 1497–1511. IEEE, 2020.

[39] Brendan Dolan-Gavitt, Patrick Hulin, Engin Kirda, Tim Leek, Andrea Mambretti, Wil Robertson, Frederick Ulrich, and Ryan Whelan. LAVA: Large-Scale Automated Vulnerability Addition. *Proceedings - 2016 IEEE Symposium on Security and Privacy, SP 2016*, pages 110–121, 2016.

[40] Andrea Fioraldi, Dominik Maier, Heiko Eißfeldt, and Marc Heuse. Afl++ : Combining incremental steps of fuzzing research. In *14th USENIX Workshop on Offensive Technologies (WOOT 20)*. USENIX Association, August 2020.

[41] Paul Fiterau-Brostean, Bengt Jonsson, Robert Merget, Joeri de Ruiter, Konstantinos Sagonas, and Juraj Somorovsky. Analysis of DTLS implementations using protocol state fuzzing. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 2523–2540. USENIX Association, August 2020.

[42] Shuitao Gan, Chao Zhang, Peng Chen, Bodong Zhao, Xiaojun Qin, Dong Wu, and Zuoning Chen. GREYONE: Data flow sensitive fuzzing. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 2577–2594. USENIX Association, August 2020.

[43] Xiang Gao, Ripon K. Saha, Mukul R. Prasad, and Abhik Roychoudhury. Fuzz testing based data augmentation to improve robustness of deep neural networks. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, ICSE '20, page 1147–1158, New York, NY, USA, 2020. Association for Computing Machinery.

[44] Heqing Huang, Peisen Yao, Rongxin Wu, Qingkai Shi, and Charles Zhang. Pangolin: Incremental hybrid fuzzing with polyhedral path abstraction. In *2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020*, pages 1613–1627. IEEE, 2020.

[45] Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts. Standard, ISO/TC 159/SC 4 Ergonomics of human-system interaction, March 2018.

[46] Kyriakos Ispoglou, Daniel Austin, Vishwath Mohan, and Mathias Payer. Fuzzgen: Automatic fuzzer generation. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 2271–2287. USENIX Association, August 2020.

[47] Zu-Ming Jiang, Jia-Ju Bai, Kangjie Lu, and Shi-Min Hu. Fuzzing error handling code using context-sensitive software fault injection. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 2595–2612. USENIX Association, August 2020.

[48] Brittany Johnson, Yoonki Song, Emerson Murphy-Hill, and Robert Bowdidge. Why don't software developers use static analysis tools to find bugs? *Proceedings of the 2013 International Conference on Software Engineering*, pages 672–681, 2013.

[49] Kyungtae Kim, Dae R. Jeong, Chung Hwan Kim, Yeongjin Jang, Insik Shin, and Byoungyoung Lee. HFL: hybrid fuzzing on the linux kernel. In *27th Annual Network and Distributed System Security Symposium, NDSS 2020, San Diego, California, USA, February 23-26, 2020*. The Internet Society, 2020.

[50] Suyoung Lee, HyungSeok Han, Sang Kil Cha, and Sooel Son. Montage: A neural network language model-guided javascript engine fuzzer. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 2613–2630. USENIX Association, August 2020.

[51] Yuwei Li, Shouling Ji, Yuan Chen, Sizhuang Liang, Wei-Han Lee, Yueyao Chen, Chenyang Lyu, Chunming Wu, Raheem Beyah, Peng Cheng, Kangjie Lu, and Ting Wang. Unifuzz: A holistic and pragmatic metrics-driven platform for evaluating fuzzers, 2020.

[52] Daniel Liew, Cristian Cadar, Alastair F. Donaldson, and J. Ryan Stinnett. Just fuzz it: Solving floating-point constraints using coverage-guided fuzzing. In *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2019, page 521–532, New York, NY, USA, 2019. Association for Computing Machinery.

[53] Baozheng Liu, Chao Zhang, Guang Gong, Yishun Zeng, Haifeng Ruan, and Jianwei Zhuge. FANS: Fuzzing android native system services via automated interface analysis. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 307–323. USENIX Association, August 2020.

[54] Valentin J. M. Manès, Soomin Kim, and Sang Kil Cha. Ankou: Guiding grey-box fuzzing towards combinatorial difference. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, ICSE '20, page 1024–1036, New York, NY, USA, 2020. Association for Computing Machinery.

[55] Valentin Manès, Marcel Boehme, and Sang Kil Cha. Fse2020 - boosting fuzzer efficiency an information-theoretic perspective, Jun 2020.

[56] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, Emanuel von Zezschwitz, and Matthew Smith. "if you want, i can store the encrypted password": A password-storage field study with freelance developers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–12, New York, NY, USA, 2019. Association for Computing Machinery.

[57] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, and Matthew Smith. Deception task design in developer password studies: Exploring a student sample. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*, pages 297–313, Baltimore, MD, August 2018. USENIX Association.

[58] Tai D. Nguyen, Long H. Pham, Jun Sun, Yun Lin, and Quang Tran Minh. Sfuzz: An efficient adaptive fuzzer for solidity smart contracts. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, ICSE '20, page 778–788, New York, NY, USA, 2020. Association for Computing Machinery.

[59] Jakob Nielsen. *Usability Engineering*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1994.

[60] Yannic Noller, Corina S. Păsăreanu, Marcel Böhme, Youcheng Sun, Hoang Lam Nguyen, and Lars Grunske. Hydiff: Hybrid differential software analysis. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, ICSE '20, page 1273–1285, New York, NY, USA, 2020. Association for Computing Machinery.

[61] Oleksii Oleksenko, Bohdan Trach, Mark Silberstein, and Christof Fetzer. Specfuzz: Bringing spectre-type vulnerabilities to the surface. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 1481–1498. USENIX Association, August 2020.

[62] Mitchell Olsthoorn, Arie van Deursen, and Annibale Panichella. Generating highly-structured input data by combining search-based testing and grammar-based fuzzing.

[63] Sebastian Österlund, Kaveh Razavi, Herbert Bos, and Cristiano Giuffrida. Parmesan: Sanitizer-guided greybox fuzzing. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 2289–2306. USENIX Association, August 2020.

[64] Soyeon Park, Wen Xu, Insu Yun, Daehee Jang, and Taesoo Kim. Fuzzing javascript engines with aspect-preserving mutation. In *2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020*, pages 1629–1642. IEEE, 2020.

[65] Hui Peng and Mathias Payer. Usbfuzz: A framework for fuzzing USB drivers by device emulation. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 2559–2575. USENIX Association, August 2020.

[66] Theofilos Petsios, Jason Zhao, Angelos D. Keromytis, and Suman Jana. Slowfuzz: Automated domain-independent detection of algorithmic complexity vulnerabilities. *CoRR*, abs/1708.08437, 2017.

[67] Jan Ruge, Jiska Classen, Francesco Gringoli, and Matthias Hollick. Frankenstein: Advanced wireless fuzzing to exploit new bluetooth escalation targets. In Srdjan Capkun and Franziska Roesner, editors, *29th USENIX Security Symposium, USENIX Security 2020, August 12-14, 2020*, pages 19–36. USENIX Association, 2020.

[68] C. Sadowski, J. Van Gogh, C. Jaspan, E. Soderberg, and C. Winter. Tricorder: Building a program analysis ecosystem. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, volume 1, pages 598–608, 2015.

[69] Sergej Schumilo, Cornelius Aschermann, Ali Abbasi, Simon Wörner, and Thorsten Holz. HYPER-CUBE: high-dimensional hypervisor fuzzing. In *27th Annual Network and Distributed System Security Symposium, NDSS 2020, San Diego, California, USA, February 23-26, 2020*. The Internet Society, 2020.

[70] Shinichi Shiraishi, Veena Mohan, and Hemalatha Marimuthu. Test suites for benchmarks of static analysis tools. *2015 IEEE International Symposium on Software Reliability Engineering Workshops, ISSREW 2015*, (November):12–15, 2016.

[71] Dag Sjøberg, Bente Anda, Erik Arisholm, Tore Dybå, Magne Jørgensen, Amela Karahasanovic, Espen Koren, and Marek Vokác. Conducting realistic experiments in software engineering. pages 17 – 26, 02 2002.

[72] Justin Smith, Lisa Nguyen Quang Do, and Emerson Murphy-Hill. Why can't johnny fix vulnerabilities: A usability evaluation of static analysis tools for security. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pages 221–238. USENIX Association, August 2020.

[73] Justin Smith, Brittany Johnson, Emerson Murphy-Hill, Bei-Tseng Chu, and Heather Richter. How developers diagnose potential security vulnerabilities with a static analysis tool. *IEEE Transactions on Software Engineering*, PP:1–1, 02 2018.

[74] Justin Smith, Brittany Johnson, Emerson Murphy-Hill, Bill Chu, and Heather Richter Lipford. Questions developers ask while diagnosing potential security vulnerabilities with static analysis. In *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*, ESEC/FSE 2015, page 248–259, New York, NY, USA, 2015. Association for Computing Machinery.

[75] David R. Thomas. A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, pages 237–246.

[76] Daniel Votipka, Michelle L Mazurek, Hongyi Hu, and Bryan Eastes. Toward a Field Study on the Impact of Hacking Competitions on Secure Development. 2018.

[77] Daniel Votipka, Rock Stevens, Elissa Redmiles, Jeremy Hu, and Michelle Mazurek. Hackers vs. Testers: A Comparison of Software Vulnerability Discovery Processes. *Proceedings - IEEE Symposium on Security and Privacy*, 2018-May:374–391, 2018.

[78] Haijun Wang, Xiaofei Xie, Yi Li, Cheng Wen, Yuekang Li, Yang Liu, Shengchao Qin, Hongxu Chen, and Yulei Sui. Typestate-guided fuzzer for discovering use-after-free vulnerabilities. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, ICSE '20, page 999–1010, New York, NY, USA, 2020. Association for Computing Machinery.

[79] Yanhao Wang, Xiangkun Jia, Yuwei Liu, Kyle Zeng, Tiffany Bao, Dinghao Wu, and Purui Su. Not all coverage measurements are equal: Fuzzing by coverage accounting for input prioritization. In *27th Annual Network and Distributed System Security Symposium, NDSS 2020, San Diego, California, USA, February 23-26, 2020*. The Internet Society, 2020.

[80] Cheng Wen, Haijun Wang, Yuekang Li, Shengchao Qin, Yang Liu, Zhiwu Xu, Hongxu Chen, Xiaofei Xie, Geguang Pu, and Ting Liu. Memlock: Memory usage guided fuzzing. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, ICSE '20, page 765–777, New York, NY, USA, 2020. Association for Computing Machinery.

[81] Valentin Wüstholz and Maria Christakis. Targeted grey-box fuzzing with static lookahead analysis. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, ICSE '20, page 789–800, New York, NY, USA, 2020. Association for Computing Machinery.

[82] Meng Xu, Sanidhya Kashyap, Hanqing Zhao, and Taesoo Kim. Krace: Data race fuzzing for kernel file systems. In *2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020*, pages 1643–1660. IEEE, 2020.

[83] Tai Yue, Pengfei Wang, Yong Tang, Enze Wang, Bo Yu, Kai Lu, and Xu Zhou. Ecofuzz: Adaptive energy-saving greybox fuzzing as a variant of the adversarial multi-armed bandit. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 2307–2324. USENIX Association, August 2020.

[84] Qian Zhang, Jiyuan Wang, Muhammad Ali Gulzar, Rohan Padhye, and Miryung Kim. Bigfuzz: Efficient fuzz testing for data analytics using framework abstraction. 2020.

[85] Rui Zhong, Yongheng Chen, Hong Hu, Hangfan Zhang, Wenke Lee, and Dinghao Wu. Squirrel: Testing database management systems with language validity and coverage feedback, 2020.

[86] Chijin Zhou, Mingzhe Wang, Jie Liang, Zhe Liu, and Yu Jiang. Zeror: Speed up fuzzing with coverage-sensitive tracing and scheduling.

[87] Peiyuan Zong, Tao Lv, Dawei Wang, Zizhuang Deng, Ruigang Liang, and Kai Chen. Fuzzguard: Filtering out unreachable inputs in directed grey-box fuzzing through deep learning. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 2255–2269. USENIX Association, August 2020.

A Semi-Structured Interview

A.1 CS Study

Task 1

- Please explain what you did in the first task.
  - Do you have a point where you want to elaborate on?
  - Did you encounter any problems?
  - Did anything went exceptionally well?
  - Please elaborate on the output of the tool.
  - Can you tell me something about the usability?
  - Where do you see potential for improvement?

Task 2

- Please explain what you did in the second task.
  - Do you have a point where you want to elaborate on?
  - Did you encounter any problems?
  - Did anything went exceptionally well?
  - Please elaborate on the output of the tool.
  - Can you tell me something about the usability?
  - Where do you see potential for improvement?

Comparison

- Please compare the two tasks.
- Do you have anything particular in mind that was comparably easy or hard?
- Would you want to use one of the tools, both or none in the future? Why?

A.2 CTF Study

Static

- Please explain what you did in the task.
- How would you rate the usability of the Clang Static Analyzer on a scale from 1-7, 1 very low, 7 very high?
- Please elaborate on the Usability of the Clang Static Analyzer.
- Can you tell me something about the Output of the analyzer?
- What was your biggest problem?
- How would you rate the documentation again on scale from 1-7?

Dynamic

- Please explain what you did in the task.
- How would you rate the usability of libFuzzer on a scale from 1-7, 1 very low, 7 very high?
- Please elaborate on the Usability of libFuzzer.
- Please elaborate on your fuzz target.
- Have you used a dictionary or corpus?
- What did you think of the output?
- How did you interact with the output?
- How did you determine that the fuzzer is running well?

Comparison

- Please compare the two tasks.
- Do you have anything particular in mind that was comparably easy or hard?
- Would you want to use one of the tools, both or none in the future? Why?

general

- What is a security related bug?

B Clang Static Analyzer Overview

Program	Clang Static Analyzer reports
Tesseract	476
protobuf 3.9.x	92
protobuf 3.8.x	121
util-linux	142
simple-obfs	15
cmatrix	3
vlc	219
wine	4746
netdata	32
darknet	73
libnice	3
obs-studio	456
jq	4
FFmpeg	639
yuzu	339
spdlog	0
simdjson	2

Table 5: Overview of GitHub projects and reports of Clang Static Analyzer

C Overview of Task Ordering

first	started	drop-out	submitted	success
Static-easy	8	1	7	6
Static-hard	7	4	3	0
second	started	drop-out	submitted	success
Static-easy	4	2	2	2
Static-hard	3	1	2	0
combined	started	drop-out	submitted	success
Static-easy	12	3	9	8
Static-hard	10	5	5	0

Table 6: CS: static analysis overall statistics

first	started	drop-out	submitted	success
Fuzzing-easy	9	5	4	1
Fuzzing-hard	7	4	3	0
second	started	drop-out	submitted	success
Fuzzing-easy	7	3	4	1
Fuzzing-hard	3	2	1	0
combined	started	drop-out	submitted	success
Fuzzing-easy	16	8	8	2
Fuzzing-hard	10	6	4	0

Table 7: CS: fuzzing overall statistics

first	started	drop-out	submitted	success
Fuzzing	3	0	3	0
Static	3	0	3	0
second	started	drop-out	submitted	success
Fuzzing	3	0	3	0
Static	3	0	3	1
combined	started	drop-out	submitted	success
Fuzzing	6	0	6	0
Static	6	0	6	1

Table 8: CTF: static analysis and fuzzing overall statistics

Participant	Static				Dynamic								
	Step 1	Step 2	Step 3	Bug	Step 1	Step 2	Step 3	Step 4	Step 5	Bug	Step 6	Corpus	Toy
CS27-SFE		no submission						not started					
CS31-SFE	✗	✓	✓	✓	✗ / ✗								✓
CS1-SFE	✓	✓	✓	✓				no submission					
CS9-SFE	✓	✓	✓	✓				no submission					
CS19-SFE	✓	✓	✓	✓				no submission					
CS15-SFE	✓	✓	✓	✓	2 / ✓	2	✗ / ✗ (FT in TPr)	✗					✗
CS5-SFE	✓	✓	✓	✓	1 / ✓	1	✓ / ✗	✓	✓	✓	✗	✓	✗
CS23-SFE	✓	✓	✓	✓	1 / ✓	1	✓ / ✓	✓	✓	✓	✓	✓	✗
CS21-SFH		no submission						not started					
CS25-SFH		no submission						not started					
CS29-SFH		no submission						not started					
CS7-SFH	✓	✓	✓	✗				no submission					
CS13-SFH	✓	✓	✓	✗				no submission					
CS17-SFH	✓	✓	✓	✗	✗ / ✗								✗
CS3-SFH	✓	✓	✓	✗	❍ / ❍	❍	✗ / ✗	✗					✓

Participant	Dynamic									Static			
	Step 1	Step 2	Step 3	Step 4	Step 5	Bug	Step 6	Corpus	Toy	Step 1	Step 2	Step 3	Bug
CS2-FSE				no submission							not started		
CS10-FSE				no submission							not started		
CS12-FSE				no submission							not started		
CS20-FSE				no submission							not started		
CS32-FSE				no submission							not started		
CS11-FSH				no submission							not started		
CS14-FSH				no submission							not started		
CS22-FSH				no submission							not started		
CS28-FSE	✗ / ✗										no submission		
CS16-FSE	✗ / ✗								✗	✓	✓	✓	✓
CS24-FSE	1 / ✓	1	✓ / ✗	✗					✓	✓	✓	✓	✓
CS6-FSE	2 / ✓	2	✓ / ✗	✓	✓	✓	✓	✗	✗		no submission		
CS18-FSH	✗ / ✗								✗	✗			
CS26-FSH	❍ / ❍	❍		❍ / ❍					✓	✓	✓	✓	✗
CS4-FSH	✗ / ✗								✓	✓	✓	✓	✗
CS8-FSH	1 / ✗	❍	✗ / ✗						✓	✓	✓	✓	✗
CS30-FSH	✗ / ✗		✗ / ❍								not started		

Table 9: CS overall

Participant	Static				Dynamic								
	Step 1	Step 2	Step 3	Bug	Step 1	Step 2	Step 3	Step 4	Step 5	Bug	Step 6	Corpus	Toy
CTF1-SF	✓	✓	✓	✗				no submission					
CTF5-SF	✓	✓	✓	✗	1 / ✓	1	✗ / ✗	✗					✗
CTF3-SF	✓	✓	✓	✗	1 / ✓	3+	✓ / ✓ (FT in TPr)	✓	✓	✗	✓	✓	✗
CTF7-SF	✓	✓	✓	✗	1 / ✓	11	✓ / ✓	✓	✓	✗	✓	✓	✗

Participant	Dynamic									Static			
	Step 1	Step 2	Step 3	Step 4	Step 5	Bug	Step 6	Corpus	Toy	Step 1	Step 2	Step 3	Bug
CTF8-FS				no submission							not started		
CTF4-FS	1 / ✓	1	✓ / ✓ (FT in TPr)	❍	❍	✗	✓		✗	✓	✓	✓	✗
CTF2-FS	1 / ✓	1	✓ / ✓ (FT in TPr)	✓	✓	✗	✓	✓	✗	✓	✓	✓	✗
CTF6-FS	1 / ✓	10	✓ / ✓ (FT in TPr)	✓	✓	✗	❍	✗	✗	✓	✓	✓	✓

Table 10: CTF overall

D Comments and Usage in Future

Participant	Comment				Use in Future		
	static		dynamic				
	positive	negative	positive	negative	static	dynamic	none
CS5-SFE	4	2	2	9	✓	✓	
CS15-SFE	4	4	1	7	✓	✓	
CS16-FSE	6	1	0	5	✓		
CS23-SFE	3	3	1	9	✓		
CS24-FSE	8	7	3	3	❍	❍	❍
CS31-FSE	2	4	1	2			✓
∑ easy	27	21	8	35	4	2	1
CS3-SFH	8	6	0	5	✓		
CS4-FSH	6	6	2	6	✓	✓	
CS8-FSH	6	6	0	7	✓		
∑ hard	20	18	2	18	3	1	0
∑	47	39	10	53	7	3	1

Table 11: CS: Comments and usage in future of the static and dynamic analysis tools

Participant	Comment				Use in Future		
	static		dynamic				
	positive	negative	positive	negative	static	dynamic	none
CTF2-FS	2	3	2	10	✓	✓	
CTF3-SF	2	5	1	2	✓	✓	
CTF4-FS	2	5	1	6		✓	
CTF5-SF	4	2	0	4	✓	✓	
CTF6-FS	2	4	2	5	clean code	✓	
CTF7-SF	8	5	0	4		✓	
∑	20	24	6	31	3	6	0

Table 12: CTF Comments and usage in future of the static and dynamic analysis tools

E Demographics

Gender	Male: 26	Female: 5	Other: 0	No Answer: 1
Age	min: 22, max: 34	mean: 26.03, median: 25	sd=2.95, NA=0	

Table 13: CS Participant Demographics

Gender	Male: 8	Female: 0	Other: 0
Age	min: 19, max: 32	mean: 23.25, median: 22	sd=4.2, NA=0

Table 14: CTF Participant Demographics

Deciding on Personalized Ads: Nudging Developers About User Privacy

Mohammad Tahaei
School of Informatics
University of Edinburgh

Alisa Frik
ICSI
University of California, Berkeley

Kami Vaniea
School of Informatics
University of Edinburgh

Abstract

Mobile advertising networks present personalized advertisements to developers as a way to increase revenue. These types of ads use data about users to select potentially more relevant content. However, choice framing also impacts app developers' decisions which in turn impacts their users' privacy. Currently, ad networks provide choices in developer-facing dashboards that control the types of information collected by the ad network as well as how users will be asked for consent. Framing and nudging have been shown to impact users' choices about privacy, we anticipate that they have a similar impact on choices made by developers. We conducted a survey-based online experiment with 400 participants with experience in mobile app development. Across six conditions, we varied the choice framing of options around ad personalization. Participants in the condition where privacy consequences of ads personalization are highlighted in the options are significantly (11.06 times) more likely to choose non-personalized ads compared to participants in the Control condition with no information about privacy. Participants' choice of ad type is driven by impact on revenue, user privacy, and relevance to users. Our findings suggest that developers are impacted by interfaces and need transparent options.

1 Introduction

Mobile advertising networks play an intermediary role of matching advertisers (companies that want to advertise their products) with publishers (apps that want to generate revenue by hosting advertising). They are a popular monetisation approach [11, 47, 57, 93, 107], with about 77% of free Android apps containing an ad library [48, 51]. To show personalized ads, ad networks collect data from app users, which raises privacy concerns [41, 111, 116]. Targeted ads can also seem intrusive and discriminating to some users [63, 83, 88, 117]. Major operating systems give users an option to limit these ads and associated tracking. However, behavioral research shows that due to status quo bias, people rarely change the default configurations [3, 52, 87, 91], and poor usability makes it hard for users to opt out of behavioral advertising and tracking [45, 56, 90]. Thus, developers' decisions regarding the defaults for their apps have implications for user privacy. Specifically, when configuring ad networks, developers can choose in the developer dashboard between personalized and non-personalized ads. Here again, status quo bias may not play out in favor of user privacy: if ad networks set personalized ads that imply more extensive personal data collection as default choices, it might nudge developers to stick to those privacy-unfriendly defaults [33, 68].

With about 24 million software developers (estimated to go up to 28.7 million by 2024) [82], who are in charge of building apps for personal smart devices, cars, and large industries, it is essential to understand how services they use may impact their decisions. Indeed, studies of privacy-related questions on Stack Overflow [106] and Reddit Android forums [59] show that developers' privacy concerns are heavily driven by large platforms such as Google and Apple. Moreover, there is a growing use of dark patterns that persuade users into make decisions that are in favor of platforms; for example, by using preselected default options, or sneaking a small product or service into users shopping basket without informing users, such as adding travel insurance during the plane ticket purchasing [42, 65, 79]. The use of dark patterns in the context of software development may have negative implications for users, as developers' choices will effect all users of their apps. For example, collecting location data, showing unrestricted ads categories, and displaying personalized ads are often allowed by default in popular ad networks [68, 102].

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

Similarly, given that ads tailored to users' preferences have a higher value [64], ad networks have incentive to nudge developers into choosing personalized ads over non-personalized ones, without necessarily acknowledging the trade-offs between revenue, user privacy, and experience. In addition to status quo bias leveraged by default choices, salience effect can be leveraged to further facilitate the nudging [18, 92]. For example, while an emphasis on user privacy may steer developers' decisions towards non-personalized ads, an emphasis on potentially larger revenue may nudge developers to choose personalized ads which is used by some ad networks through including statements like "including personalized ads may likely result in higher revenue" in their documentation, quick start guides, and blog posts [102, 104, 108].

In this study, we aim to understand how choice framing in ad networks effects developers' decision making. Our research question are:

RQ1: How does choice framing in ad networks impact developers' decisions about ad personalization?

RQ2: What are the reasons behind developers' choices of personalized or non-personalized ads?

To answer our research questions, we conducted an online survey-based experiment with 400 participants with app development experience. In a hypothetical scenario, we asked them to make a series of choices to integrate ads in a personal finance management app and a gaming app. The main decision of interest was regarding the choice between personalized and non-personalized ads. The framing of those choices was manipulated between one control and five experimental conditions, to emphasize implications for framing around data processing restrictions, user-facing descriptions, user privacy, developer's revenue, and both user privacy and developer's revenue. To help further contextualize and interpret the results, we also surveyed participants' opinions and attitudes about personalized ads, ad networks, and privacy regulations.

We find that although on average the majority of participants decided to integrate the personalized ads, choice framing significantly impacted their decisions. When user privacy implications were made salient, participants were 11.06 times more likely to select non-personalized ads than when the neutral framing was used (Control condition). When a framing emphasized data processing restrictions, participants were 3.45 more likely to select the non-personalized ads than in the Control condition. Other nudges—emphasizing the consequences of ads on an app's revenue, presenting participants with an explicit choice between user privacy and app's revenue, and telling participants that users will be able to see whether the app is using ads based on their personal data or not—did not significantly change participants decisions compared to the Control condition.

The analysis of open-ended responses revealed a variety of reasons for developers' choices, ranging from maximizing the app's revenue and relevance of ads to the uses, to concerns about user privacy and regulation compliance, and implications for user experience. From the exit survey, we found that even when upper and middle management choose the ad networks and app's business models, developers still feel involved in this decision-making process. However, developers generally believe that they do not have full control over ad networks' data collection, and believe users have even less control. By illustrating the potential impact of choice framing on ad personalization decisions during app development, our results inform regulators about the need to enforce greater control over ad networks' data collection and analysis practices, discourage from using dark patterns, and encourage ad networks to adopt interfaces for developers that may assist them in making informed decisions about user privacy.

2 Related Work

Ad Networks. Ad networks are a popular mobile app monetisation approach [11, 47, 57, 93, 107]. Over half of Android apps include ad network libraries [11, 48, 51, 107], which often offer both personalized and non-personalized ads. Personalized ads attract more user attention than non-personalized ads [20, 63], generating higher engagement and therefore revenue. To provide ads tailored to a specific user, ad networks collect personal information from users such as age, gender, and location [84, 100], not only in free apps that rely mostly on ads to generate revenue, but also in paid apps [19, 47]. However, personalized ads have some negative consequences for users. For example, some users find them discomforting [63, 117], discriminating [86], and intrusive [83, 88].

Options Provided by Ad Networks to Users and Developers. Both users and developers can limit data collection and turn off ad personalization. After the introduction of the General Data Protection Regulation (GDPR) [39] and the California Consumer Privacy Act (CCPA) [26], the prevalence of these options particularly increased [50].

On the user side, self-regulatory programs (e.g., Digital Advertising Alliance opt-out [31]), smartphone operating systems, service providers, and browsers offer settings that allow opting out of ad personalization [66], and at minimum, request user consent to show personalized ads. Research shows limited effectiveness, usefulness, legal compliance [37, 46, 67, 112], and usability [67, 81] of these methods.

On the developer side, ad networks provide an interface for configuring personalization and data collection for specific apps and geographic regions. These interfaces often use defaults that are not in favor of user privacy [68, 102]. Developers tend to keep the defaults, follow industry standards, guidelines, and requirements provided by the platforms built by large tech companies [43, 59, 95, 106] without fully considering all the options and consequences of their choices on user privacy [30, 33, 68]. Developers generally acknowledge the value of user privacy [33, 68, 94], but find it challenging

to understand what information is collected, how it is used by platforms [33, 68, 103], and how to protect user privacy [59, 106]. Hence, some poor user privacy elements in how apps integrate ad networks may be caused by the way ad networks are framing choices and nudging developers through defaults.

Nudging. Humans can be nudged towards making certain actions through the use of specific wordings, framing, colors, and default values [3, 27]. *Choice framing*, in particular, uses the activation of salience effects [18, 92] and status quo bias [52, 87, 91], to effectively nudge the privacy choices of users [3, 16]. For example, priming survey respondents about privacy using words like "privacy-sensitive" and "potential privacy risks" increases the reported privacy concerns [25] and making privacy information salient drives more privacy-preserving choices in user experiments [109]. We believe that similar effects can be achieved in the context of software development, where choice framing in tools and interfaces may affect developers' decision making.

Nudges can be used to encourage users to make decisions that are favorable to service providers (e.g., ad networks) but not necessarily favorable to themselves. Such practices are often referred to *dark patterns*—"instances where designers use their knowledge of human behavior (e.g., psychology) and the desires of users to implement deceptive functionality that is not in the user's best interest" [42, p. 1]. In the context of privacy, the examples of dark patterns include privacy consent forms that do not provide a "reject all" button [81] and hard-to-find (or completely absent) options for deleting accounts [23]. Similar patterns are also visible in ad networks' developer dashboards where the default values are all set to personalized ads and location data is often collected by default [68, 102].

Our Contribution. We extend the literature on developer-facing privacy interfaces by looking at the privacy nudges directed at developers and exploring the impact of choice framing in ad networks' developer dashboards.

3 Method

To answer our research questions, we conducted an online survey-based between-subject experiment with 400 participants with mobile development experience administered using Qualtrics. The study received ethical approval from our institute. All participants provided informed consent before completing the study. We describe the study protocol below, and the full survey text is in Appendix A.2.

After screening for app development experience (Section 3.2), participants were randomly assigned to one of six conditions (Section 3.1), and asked to complete the main survey. Each participant was presented with two hypothetical scenarios in a random order: one was about a *gaming app*, another one was about a *financial app* for personal finance management. We chose these app categories, because personal finance management has obvious privacy implications (e.g., developers reported more sensitive variables for the financial category compared to other app categories [17]), and gaming is the most popular category on both Apple App Store and Google Play [76, 77].

Participants were asked to imagine that they were a shareholder in a software development company, and together with a small team, they created a (financial or gaming) app, which will be published in Europe and the United States and is mainly targeted towards adults above the age of 18. Then, we asked them to answer questions posed by the "Acme Assistant", a tool for an imaginary ad network that helps with integrating the ad network into the app. The Assistant was inspired by MoPub Integration Suite, a new service by Twitter's MoPub ad network for an easy app integration [74]. The Assistant asked five multiple-choice questions about ad formats (e.g., banner and interstitial), level of graphics (high-quality and moderate-quality), platforms (e.g., Android and iOS), types of ads (personalized and non-personalized), and the regulations that apply to the app (e.g., GDPR, CCPA). After making the choices, they were also asked an open-ended question about the primary reason for choosing the personalized or non-personalized ad type.

After completing the above for both the financial and gaming apps, they were sent to an exit survey with the questions about: how they would go about asking for user consent for the personalized ads, how the choice of ad type would affect an app's revenue or number of users, what role does user privacy play in their daily development routines, and how much users and developers have control over data collected by ad networks. The exit survey provided additional insights about participants' opinions, knowledge, and attitudes, and helped to further contextualize and interpret experimental results. Finally, they answered software and mobile development, and demographics questions.

3.1 Experimental Conditions

All participants were randomly assigned to one of six conditions including one Control group and five treatment groups. The only difference among the conditions was the framing of the choice about personalized or non-personalized ads. The order of all options was randomized. Each choice consisted of a short label phrase followed by a longer description.

Control–Minimal Information ($N = 66$): *(1) Personalized ads: Acme can show personalized ads to your users. (2) Non-personalized ads: Acme will show only non-personalized ads to your users.* This framing was inspired by Google AdMob's developer dashboard to help developers build GDPR-compliant apps for European users (Figure 2 in the Appendix). It used neutral wording about ad types without mentioning any information about collection and processing of user data.

Data Processing Restrictions ($N = 67$): *(1) Ads with unrestricted data processing: Acme can show personalized ads to your users based on a user's past behavior, such as previous visits to sites or apps or where the user has been. (2) Ads with restricted data processing: Acme will show only non-personalized ads to your users based on contextual information, such as the content of your site or app, restricting the use of certain unique identifiers and other data.* This framing was inspired by Google AdMob's developer dashboard to help developers build CCPA-compliant apps for California users (Figure 3 in the Appendix) and it explicitly hinted at the types of data used for ad personalization, which may indirectly encouraged developers to consider privacy implications of such data processing. We based two of our conditions on Google AdMob because it is the most common mobile ad network in apps [6, 7, 40].

User-Facing Descriptions ($N = 68$): *(1) Ads with 'Personalized Ads' tag displayed to users: Acme can show personalized ads to your users. Users will see the 'Personalized Ads' tag next to the 'Install' button and the following text in your app description in the App Store or Google play "This app shows ads personalized based on your personal information." (2) Ads with 'Non-personalized Ads' tag displayed to users: Acme will show only non-personalized ads to your users. Users will see the 'Non-personalized Ads' tag next to the 'Install' button and the following text in your app description in the App Store or Google play "This app shows ads not personalized based on your personal information."* This condition aimed at leveraging transparency and nudging developers' accountability and responsibility to users. The framing was inspired by the recent additions to the Apple App Store called "Privacy Details" to "help users better understand an app's privacy practices before they download the app on any Apple platform" [12] and prior work's recommendation about including privacy features of apps in the app stores to softly nudge developers to consider user privacy in their apps [59].

Privacy Focused ($N = 67$): *(1) Ads with lower user privacy: Acme can show personalized ads to your users based on their past behavior, such as previous visits to sites or apps or where the user has been. (2) Ads with higher user privacy: Acme will show only non-personalized ads to your users based on contextual information, such as the content of your site or app.* This condition is aimed at leveraging salience effects [18, 92], by making privacy implications prominent in the choice option descriptions.

Revenue Focused ($N = 65$): *(1) Ads with higher revenue: Acme can show personalized ads to your users, which may yield higher revenue than non-personalized ads. (2) Ads with lower revenue: Acme will show only non-personalized ads to your users, which may yield lower revenue than personalized ads.* This condition aimed at leveraging salience effects [18, 92], by making revenue implications prominent in the choice option descriptions.

Privacy vs. Revenue ($N = 67$): *(1) Ads with higher revenue: Acme can show personalized ads to your users, which may yield higher revenue than non-personalized ads. (2) Ads with higher user privacy: Acme will show only non-personalized ads to your users which may increase your users' privacy.* This condition aimed at exploring what choices the participants would make if they were faced with an explicit trade-off between the user privacy and revenue.

3.2 Recruitment and Screening

We used Prolific, GitHub, and LinkedIn groups to recruit the participants (Jan '21). On average, the survey took 19 minutes ($SD = 89$, $median = 13$) to complete. The large standard deviation was due to some participants who left the survey open but stepped away before returning and completing it.

Prolific. Using Prolific's exclusion criteria, we recruited 1,288 participants who were fluent in English, had computer programming skills, and an approval rate of at least 90%. They responded to a 1-minute screening survey (Appendix A.1) to assess their software development experience, and received £0.15 compensation. Those who worked on at least one app in the past three years ($N = 466$) were invited to the main survey and were paid £1.50 for completing it. Of the invited participants, 372 respondents started the main survey, but eight did not complete it. We removed two respondents because they had worked on over eighty apps while having less than three years of mobile development experience, one respondent who finished the survey in less than three minutes, and one respondent who did not pass the attention check question. In total, we received 328 valid responses from Prolific.

GitHub. We sent emails to GitHub users who contributed to the top 1,000 GitHub repositories (sorted by the number of stars) written either in (1) Java (with "Android" as an additional keyword), or (2) Objective-C or Swift (with "iOS" as an additional keyword). In total, we sent out 33,675 emails, out of which 128 started the survey, 51 respondents did not finish the survey, and five had not developed apps in the past three years. Other checks did not result in removing any additional responses. In total, we received 72 valid responses from GitHub emails. These participants were offered to provide an email to enter into a raffle for a £30 gift card for each 20 participants; 57 participants decided to enter the raffle, out of which three random participants received a gift card.

Other Channels. We made an effort to recruit women and minority groups by posting the survey in 20 LinkedIn groups specific to these populations. 14 respondents started the survey, seven did not finish the survey, and the other seven had not worked on any apps in the past three years. Therefore, we did not receive any valid responses from these channels.

The anonymized dataset for multiple-choice responses, excluding the open-ended responses (per participant consent), for the 400 valid participants is available online at DOI: 10.7488/ds/3045.

3.3 Data Analysis

3.3.1 Quantitative Analysis

We fitted a generalized linear mixed model with the binary value of choice between personalized (coded as 0) and non-personalized ads (coded as 1) as the dependent variable because each participant contributed two output values, one per app category. The model consisted of the six conditions (with Control as the baseline), app category (with gaming as the baseline), and several demographics as fixed effects, and participants as random effects, given that we had two data points per participant (gaming and financial apps) [73]. The regression analysis was conducted in R using the lme4 (`glmer`) [21] and arm [38] packages using binomial family (`logit` was the link function).

3.3.2 Qualitative Analysis

The count of words in the three open-ended questions showed that the answers were brief (on average 20 words, $SD = 16$) and enabled us to use affinity diagrams—a tool for organizing and consolidating output from a brainstorming session according to its affinity, or similarity—for analysis [24, 55]. We used the virtual collaboration platform Miro [72] to create separate boards for each open-ended question and posted virtual sticky notes with participants' responses. During a half-day virtual session with five security and privacy researchers with a minimum Master's degree in computer science, and one senior Android developer, we identified the common themes through group affinity diagram building.

3.4 Limitations

As with any self-reported data, respondents' survey answers may be subject to social desirability bias [36] and may differ from actual behaviors (so called, privacy paradox [54]). However, our use of role-playing scenarios and questions about intentions (rather than only attitudes) partially mitigates these biases, as intentions are shown to significantly correlate with behaviors [8, 32]. Our work complements and extends other privacy-related studies with developers [59, 101, 106] by conducting a controlled study with high internal validity which provides a foundation for future validation work. The results show a promising effect which will need further field experiments to fully test the generalizability.

Compared to other studies using similar recruitment strategies, the response rate for GitHub emails in our study is 0.21%, which is similar to 0.31% in [105] and lower than 1.3% in [1]. However, we were able to recruit a sufficient number of participants through Prolific. Moreover, mentioning ad networks in the recruitment email could deter people concerned about user privacy or ad networks. However, our results do not support that worry, demonstrating a wide variety of opinions about ad networks and user privacy.

Due to the demographic composition of the Prolific participant pool [35], our sample is predominately European, which could result in participants being more aware of European privacy laws, i.e. GDPR. However, GDPR's jurisdiction applies worldwide and many developers create apps for different geographic markets, mitigating this concern. To geographically balance our sample, we used additional Prolific screening criteria to exclude European countries for 274 respondents of the screening survey. The diverse geographic background of GitHub participants also added diversity to our sample. While our results may not be generalizable to all populations, it provides insights on the impact of various nudges on developers' decisions. Additionally, including geographic variable did not improve our model's fitness and did not reveal any significant relation to the outcome variable. Future research is encouraged to validate the results with other populations.

Identification of participants as developers was self-reported, as we did not test them. However, we believe it does not undermine the validity of results, as GitHub is a platform targeted at developers, and Prolific participants had previously marked themselves as having computer programming experience. The recruitment materials also highlighted that the study was about improving advertising library integration experience; such jargon is likely to defer participants without relevant experience and attract developers.

4 Results

We first report participants' demographics in Section 4.1, then the main experimental effects in Sections 4.2 and 4.3, and finally the additional findings about participants' opinions and attitudes about ads personalization in Section 4.4 to contextualize and interpret the main results.

4.1 Participants

Our participants are mostly European (66%), male (82%)—representative of the male-dominated profession [98], have on average 5.1 years of experience in software development ($SD = 5.3$), 2.7 years of experience in mobile development ($SD = 2.6$), on average worked on 3.5 apps in the past three years ($SD = 4.2$), 73% worked in software teams (e.g., developer, tester, or manager), 46% hold a software development position, 69% had previously integrated an ad library, and 78% make money from software development (see Table 5 in the Appendix). Over 90% of Google Play developers have one to nine apps under their account (as of 2015) [115], suggesting that our sample represents a portion of mobile developers. More than half (57%) of participants have used at least one ad network in their apps. Google AdMob (48%), Facebook Audience Network (20%), and Unity Ads (20%) were the most popular ad networks.

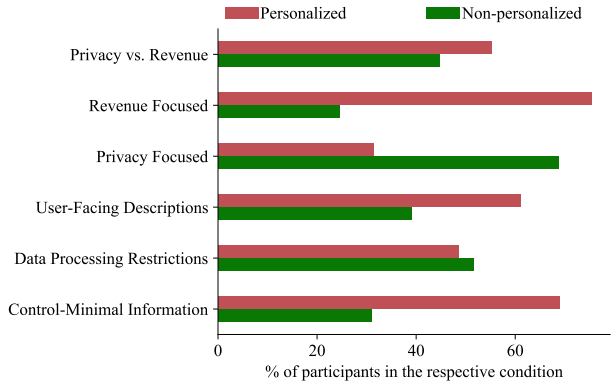


Figure 1: Participants' choices between personalized and non-personalized ads across the six conditions.

4.2 Choices Between Personalized and Non-Personalized Ads

As shown in Figure 1 (RQ1), the majority of participants chose personalized ads in the Revenue Focused (75%), Control (69%), and User-Facing Description (61%) conditions, and non-personalized ads in the Privacy Focused condition (69%). In the Data Processing Restrictions and Privacy vs. Revenue conditions, the choices between the two types of ads were split almost equally, with 49% and 55% respectively choosing the personalized ads.

The regression analysis (Table 1) confirms that the choice framing does impact participants' choices (RQ1). The strongest effect was in the Privacy Focused condition: using framing that explicitly mentions the implication for user privacy and what data will be used nudged participants to be 11.06 times ($p < .001$) more likely to choose non-personalized ads over personalized ads, compared to the Control condition. In the Data Processing Restrictions condition, framing that emphasized data restrictions associated with the choice of ads nudges participants to be 3.45 times ($p = .011$) more likely to choose the non-personalized ads compared to the Control condition. The results in the Revenue Focused, User-Facing Descriptions, and Privacy vs. Revenue conditions were not significantly different from the Control condition. In other words, using the neutral framing about personalized and non-personalized ads (Control condition), emphasizing the consequences of personalized ads on app's revenue (Revenue Focused condition), leveraging the user-facing description to provide transparency to users about whether app uses personalized ads based on users' personal data or not (User-Facing Description condition), and providing an explicit choice between user privacy and app's revenue (Privacy vs. Revenue) similarly affect participants' choices to integrate predominantly personalized ads in the apps.

Impact of App Category: Financial vs Gaming. Participants' choices between the app categories were not different; 57% of participants chose personalized ads in both categories. Thus, our expectation that the financial app would trigger more privacy-preserving choices (non-personalized ads) because it carries obvious privacy risks for users is not supported by the data. We did not observe a significant interaction between conditions and app categories. In Section 5.3, we explore the potential reasons behind this effect based on participants' open-ended answers.

Independent Variables	**ORs**	**CI (95%)**	***p*–value**
Condition			
Control–Minimal Information		Reference	
Data Processing Restrictions	3.45	1.32–8.98	.011*
User-Facing Descriptions	1.38	0.54–3.50	.502
Privacy Focused	11.06	3.97–30.75	<.001***
Revenue Focused	0.50	0.19–1.33	.164
Privacy vs. Revenue	2.48	0.97–6.35	.058
App Category			
Gaming app		Reference	
Financial app	1.02	0.70–1.49	.923
Given Priority to Privacy in Development Routines			
Low priority		Reference	
Not a priority	1.27	0.11–15.04	.851
Medium priority	1.84	0.75–4.51	.184
High priority	3.94	1.59–9.75	.003**
Essential	10.33	3.43–31.11	<.001***
Main Income Source			
Salary, not dependent on app revenue		Reference	
Don't make money from app development	2.63	1.23–5.66	.013*
Salary, partially dependent on app revenue	0.57	0.27–1.17	.126
Direct app revenue	0.73	0.32–1.66	.447
Other	0.90	0.07–11.17	.934
Years of experience in software development	1.08	1.02–1.14	.007**
Number of developed apps in the past three years	0.92	0.86–0.99	.033*
(Intercept)	0.09	0.03–0.3	< .001***

Table 1: Generalized linear mixed model regression. Outcome variable is the binary choice between personalized (coded as 0) and non-personalized ads (coded as 1). OR: odds ratios, CI: confidence intervals, conditional R^2: .614 (represents how much of the variance is explained by the model [62]), No. observations: 800, $*p < .05$, $**p < .01$, $***p < .001$.

Impact of Demographics. We also included the demographic variables in the model that improved the model's fit. We found that participants, who consider privacy an essential or high priority are 10.33 ($p < .001$) and 3.94 times ($p = .003$), respectively, more likely to choose non-personalized ads compared to those who consider privacy a low priority in daily development routines (we selected the low priority as the reference category here because the not a priority category only had five responses making the category sizes highly unbalanced). Participants, who do not make money from software or apps, are 2.63 times ($p = .013$) more likely to choose the non-personalized ads compared to those

whose income is from software/app development but is not dependent on app revenue.

Each additional year of experience in software development increases the likelihood of choosing non-personalized ads by 8% ($p < .001$), but each additional app that participants developed in the past three years decreases the odds of choosing the non-personalized ads by 8% ($p = .033$).

The inverse relation between the number of developed apps and the choice of non-personalized ads may be related to the participants getting used to the status quo in that area as they develop more apps. More years of experience may also increase developers' awareness about other app monetisation methods. Inclusion of other variables, such as years of experience in mobile development, did not improve the model fit, thus we did not include them in the final model.

4.3 Reasons Behind the Ad Type Choices

Using affinity diagrams, as discussed in Section 3.3, we constructed themes around participants' responses to the question: "What was the biggest reason that made you pick the ad type: [their choice]" (RQ2). Table 2 shows the resulting themes. We provide the unique count of participants that mention each theme at all (out of 400) as well as the number of responses that mention a theme (out of 800) as each participant provided a response for each of the two apps. Quotes are labeled with P or NP based on the participant's choice for personalized or non-personalized ads. Theme frequencies are provided to give a sense of scale, but should not be used for generalization or statistical analysis since they only measure what participants thought to mention.

We identified three major reasons for choosing personalized or non-personalized ads: expected impact on revenue, user privacy, and relevance to users. Participants in the Privacy Focused condition mentioned privacy most often, and participants in the Revenue Focused condition mentioned monetisation most often as a reason for their ads choices.

Impact on Revenue. A main reason for choosing a certain ad type was related to monetisation goals and impact on revenue, mentioned by 41.5% of participants (166/400). Those, who chose personalized ads, were especially likely to relate their choice to expected positive impact on revenue (232/800): "To ensure most people click on the ad, increasing the apps revenue" (P309). Less often participants chose non-personalized ads with the expectations of positive impact on revenue (24/800): "I believe that providing non-customized ads would help to increase consumption regardless of the type of ad" (NP68).

User Privacy. Out of participants who chose non-personalized ads, most did it because of user privacy (269/800), for example, to protect users' sensitive data (35/800), gain their trust (40/800), comply with privacy regulations (13/800), or gain a competitive advantage (12/800): "App doesn't have personalized information about the user. Also, it is easier to comply with GDPR rules that way" (NP213), "Given Apple's latest privacy changes, users are more aware of apps that invade their privacy and as a result, could be less likely to download these apps" (NP224). Some mentioned the long-term benefits of user trust over the short-term gains from violating user privacy: "Users trust in protecting the privacy is the most valuable good for a developer (besides quality of content). Aiming at a one-hit-wonder one wouldn't care about it, but with long time plans this is the only manageable compromise for all stakeholders" (NP135).

Participants, who mentioned privacy in relation to their choice of personalized ads (24/800), mostly assumed that users do not care about privacy (7/800): "Just like it is with facebook and other big ad circulators, It's proven that people only care about their privacy on a surface level" (P202).

Several participants acknowledged the trade-off between user privacy, trust, and other considerations such as revenue (6/400): "I was torn. On the one hand, personalized ads in the context of ones [sic.] finances are going to have a *much* higher CPM and I would like to capitalize on that. However, because I'm running an app whose data is sensitive and where I am more dependent on long term trust from my users, I decided to make the ads less personalized to start so that I can have fewer scary disclosures and consent screens. If the app is successful, I can always explore personalizing them later" (NP197). Participants also expressed struggling with the trade-off between revenue and user privacy: "Desire to protect customers privacy. This was a tough one and I waffled back and forth. If it offered higher payout I would have selected this option" (NP317).

Only seven participants mentioned the potential security risks associated with personalized ads: "This type of app wants to give the user a sense of security so personalized ads might put someone off from using this app to manage their finances" (NP473).

Relevance to Users. Many participants believe that ads should be interesting, relevant, engaging, and useful to the users (156/400). On the one hand, they believe that such ads are beneficial to the users: "Personalized ads are appealing to the user, a person interested in a specific topic would rather see/read more about it than a random ad" (P169). Given that personalized ads are targeted to users' potential interests, most participants driven by that reason selected the personalized ads over non-personalized ones (197/800 vs. 29/800). A smaller group of participants chose non-personalized ads because they considered them relevant to users: "Using non-personalized ads, you have the luxury of inserting different ads of which some may get the attention of the users further increasing the interaction" (NP163). Some participants were even worried that relevant ads may distract users' attention

Theme	Condition (participants, $N = 400$)							Ad Type Choices (occurrences, $N = 800$)		
	Control	Data Processing Restrictions	User-Facing Descriptions	Privacy Focused	Revenue Focused	Privacy vs. Revenue	Total	Personalized	Non-Personalized	Total
Impact on revenue	32 (8.0%)	16 (4.0%)	29 (7.2%)	18 (4.5%)	46 (11.5%)	25 (6.2%)	166 (41.5%)	232 (29.0%)	24 (3.0%)	256 (32.0%)
User privacy	13 (3.2%)	34 (8.5%)	23 (5.8%)	48 (12.0%)	11 (2.8%)	32 (8.0%)	161 (40.2%)	24 (3.0%)	269 (33.6%)	293 (36.6%)
Sensitive data	1 (0.2%)	9 (2.2%)	4 (1.0%)	11 (2.8%)	1 (0.2%)	6 (1.5%)	32 (8.0%)	-	35 (4.4%)	35 (4.4%)
User trust	2 (0.5%)	6 (1.5%)	5 (1.2%)	3 (0.8%)	7 (1.8%)	7 (1.8%)	30 (7.5%)	5 (0.6%)	40 (5.0%)	45 (5.6%)
Compliance	-	6 (1.5%)	1 (0.2%)	2 (0.5%)	2 (0.5%)	1 (0.2%)	12 (3.0%)	3 (0.4%)	13 (1.6%)	16 (2.0%)
Competitive advantage	-	3 (0.8%)	-	3 (0.8%)	-	4 (1.0%)	10 (2.5%)	-	12 (1.5%)	12 (1.5%)
Users don't care about privacy	-	-	-	6 (1.5%)	-	1 (0.2%)	7 (1.8%)	7 (0.9%)	-	7 (0.9%)
Security reasons	1 (0.2%)	2 (0.5%)	1 (0.2%)	2 (0.5%)	-	1 (0.2%)	7 (1.8%)	1 (0.1%)	8 (1.0%)	9 (1.1%)
Privacy & ethics trade-off	-	-	1 (0.2%)	3 (0.8%)	1 (0.2%)	1 (0.2%)	6 (1.5%)	4 (0.5%)	4 (0.5%)	8 (1.0%)
Relevance to users	33 (8.2%)	26 (6.5%)	33 (8.2%)	11 (2.8%)	30 (7.5%)	23 (5.8%)	156 (39.0%)	197 (24.6%)	29 (3.6%)	226 (28.2%)
User experience	8 (2.0%)	9 (2.2%)	17 (4.2%)	12 (3.0%)	11 (2.8%)	3 (0.8%)	60 (15.0%)	48 (6.0%)	27 (3.4%)	75 (9.4%)
Category-related reasons	7 (1.8%)	9 (2.2%)	6 (1.5%)	18 (4.5%)	5 (1.2%)	15 (3.8%)	60 (15.0%)	18 (2.2%)	89 (11.1%)	107 (13.4%)
Finance-related	3 (0.8%)	9 (2.2%)	4 (1.0%)	13 (3.2%)	5 (1.2%)	8 (2.0%)	42 (10.5%)	7 (0.9%)	72 (9.0%)	79 (9.9%)
Gaming-related	3 (0.8%)	2 (0.5%)	3 (0.8%)	7 (1.8%)	-	8 (2.0%)	23 (5.8%)	10 (1.2%)	20 (2.5%)	30 (3.8%)
Specificity of a target audience	2 (0.5%)	-	1 (0.2%)	5 (1.2%)	4 (1.0%)	5 (1.2%)	17 (4.2%)	10 (1.2%)	10 (1.2%)	20 (2.5%)
Users should decide	2 (0.5%)	2 (0.5%)	5 (1.2%)	4 (1.0%)	2 (0.5%)	2 (0.5%)	17 (4.2%)	18 (2.2%)	8 (1.0%)	26 (3.2%)
Easier to develop	-	3 (0.8%)	1 (0.2%)	-	4 (1.0%)	-	8 (2.0%)	1 (0.1%)	9 (1.1%)	10 (1.2%)
Everyone does it	-	1 (0.2%)	1 (0.2%)	3 (0.8%)	1 (0.2%)	1 (0.2%)	7 (1.8%)	7 (0.9%)	-	7 (0.9%)
Unclear responses	5 (1.2%)	4 (1.0%)	2 (0.5%)	4 (1.0%)	1 (0.2%)	3 (0.8%)	19 (4.8%)	15 (1.9%)	11 (1.4%)	26 (3.2%)

Table 2: Constructed themes from participants' answers about the primary reason for choosing the ad type.

away from the app, reducing engagement: "You would get distracted if you saw a product that you like, the user could easily close the app and search that product" (NP42).

Participants in the Privacy Focused condition were least likely to mention the relevance of ads to the users (11/400), but we did not observe much difference among the other conditions (23–33/400).

User Experience. Some participants (60/400) mentioned the impact of ads on user experience as a reason for their choice. In contrast to the theme about relevance of ads emphasizing their utility and benefits to the users, this theme emphasizes the emotional and experiential impact of ads.

Participants who chose personalized ads (48/800) thought that they are less annoying, more enjoyable, and of higher quality: "To avoid frustrating customers with irrelevant to their interests ads that they will be forced to watch throw [sic.] to play the game for free personalized ads are a great choice to make fun the rewarded video ad format" (P493), ". . . I would like the ads to feel native to the app so it is a more professional experience for the user and as such high quality and personalized ads would fit better for such an app" (P333). Participants who chose non-personalized ads (27/800) believed them to be less invasive and creepy: "I feel that personalized ads are too intrusive and creepy, so I would rather opt for non-personalized ads. . . . I don't want to scare away users" (NP330). Some participants preferred to reduce the number of ads in general to minimise the interruption of the main interaction with the app, especially in the gaming context: "Gaming isn't a prime state to be in to think about purchases. As someone with experience, ads feel like a break in action in games and I would say its not worth the extra money overall" (NP396).

Category-Related. Some participants said their choice of ad type partially depends on the app category, the data it collects, or the specific user audience it targets (60/400). For instance, we already discussed earlier that perceived sensitivity of user data may raise privacy and trust concerns, especially in the context of a financial app, leading participants to choose non-personalized ads: "We're building a financial app after all. The data in there is sensitive and if there have to be ads, they should in no way track the user. Otherwise we'll loose trust faster than we can build the app" (NP136). Similarly, some participants thought that the data collected in the gaming app is not sensitive, justifying the use of personalized ads: "The information shared with a gaming type of application may be not as important to the consumer" (P301). Others thought that the data collected in the gaming app does not reveal personal information, and thus cannot be used for targeting, leading to the choice of non-personalized ads: "A Gaming app should not have any access to personal data, so personalized advertising is just not possible" (NP192).

On the other hand, a few participants (6/400) thought that the target audience of a financial app is particularly valuable to advertisers, due to their higher buying power, thus, promising a particularly high return on personalized advertising: "The target market for the app is an older and more affluent audience, therefore it is worth exploring to show the personalized ads to yield a higher revenue" (P474).

Other Themes. These themes were mentioned by a few participants, but still provide interesting insights. For instance, 17 participants said that they prefer to let *users* decide what types of ads they want to see. For example, participant P39 shifted the responsibility to users assuming that they know what information was used for customizing the ad, what the privacy implications are of such targeting, and what the appropriate tools are for controlling online tracking: "Because I bet on the smart mind of my client, he/she should know how ads work and should know whether if the ad is shown after seeing custom profiling data or not and to offer the choice to get tracked or not" (P39). Participant NP299 acknowledged that there is currently little transparency about the data practices in app stores, and that users may not pay attention to the disclosures with poor usability: "Somehow in google play they do not give at least warnings and most users install without first reading labels. The case is to leave that label so that the user reads or does not read it is aware of the type of advertising that is included with the application" (NP299).

Eight participants expected that it will be easier and faster to implement non-personalized ads: "Helps to get app on stores, we are not collecting personal information and it helps to pass faster" (NP12). Seven participants chose personalized ads simply because it is common and it is the status quo in app advertising: "Many of the apps that I use have this type of ad" (P484).

4.4 Opinions About Ad Networks, Privacy Regulations, and Consent

In this section we report the results from the exit survey that helped us further contextualize and interpret the main treatment effects, as later discussed in Section 5.

Perceived Control Over Ads. While the choices about ad networks' and apps' business models are often made by upper-level and middle management (Figure 4 in the Appendix), our participants feel involved in that decision-making process. Many participants have been involved at least a moderate amount in choosing ad networks (36%), configuring ads (46.7%), and integrating the code to enable in-app ads (47.5%) (Figure 5 in the Appendix). However, despite the involvement in selecting ad networks, participants mostly agree that developers have moderate (40.25%) or very little (32.75%) control over the data collection by those networks (Figure 6 in the Appendix); and end-users have even less control (Wilcoxon signed-rank test of perceived end-user control relative to developer control: $U = 8409$, $p < .001$).

Reasons for Not Including an Ad Network. More than half (69%) of participants have used at least one ad network in their apps. We asked the remaining 123 participants to explain why they did not include any ad networks in their apps and constructed themes around participants' answers (Table 3), as discussed in Section 3.

Reason for Not Including Ad Networks	#Participants
No need to monetize the app	50 (40.65%)
Generic reasons	31 (25.2%)
Paid apps	12 (9.8%)
Open-source or free apps	7 (5.7%)
Apps not intended for public audience	25 (20.3%)
Small and personal projects	17 (13.8%)
Academic projects	8 (6.5%)
Expected negative impact on user experience	18 (14.6%)
Decision was made by others	16 (13.0%)
It's a responsibility of others	7 (5.7%)
Don't know how to do it	5 (4.1%)
User privacy	4 (3.3%)
Still in early development stages	4 (3.3%)
Unclear responses	4 (3.3%)

Table 3: Constructed themes around participants' reasons for not including ad networks in their apps ($N = 123$).

Forty percent of these participants (50/123) did not integrate ad networks because there was no need to use ads to monetize the app, for instance, because it was free or open-source, or relied on other sources of revenue. About 20% of participants (25/123) did not aim for a broad audience and public use, but used instead for small personal projects, learning experience, homework, or academic research. Some participants (18/123) considered ads intrusive and damaging to user experience: "I've always found it less intrusive for the end-users and a much smoother experience for them overall so buying a premium version would be preferred as a way to monetize the apps" (P131). Others (16/123) said that they did not have control over that decision, e.g., because they were developing an app for a client. A few participants said that they did not know how to integrate an ad network (5/123), it was someone else's responsibility to do it (7/123), or the project was still in the early development stage for ad integration (4/123). Only four participants explicitly mentioned concerns about user privacy: "Ad networks are not transparent and can't be audited. I can't control the amount of information fetch from my users" (P201).

Perceived Impact of Personalized Ads on Revenue and User Base. We asked participants how choosing personalized ads over non-personalized ads is likely to affect the revenue and number of users (Figure 7). The majority of participants expected an increase in revenue in both app categories, but no or little decrease in the user base. Specifically, almost half of participants expected an increase in revenue by

up to 40%. Slightly more participants believed that the user base won't change in the gaming app compared to financial app (43% vs. 32.5%). However, 16-18% of participants believed that deploying personalized ads will not change the revenue at all, or even *decrease* the revenue in both app categories, and decrease the user base by up to 40% in financial (32%) and gaming (23%) apps.

Beliefs About Privacy Regulations. In the survey scenarios, we told participants that the apps will be published in Europe and the United States and are mainly targeted towards adults above age of 18. For both apps, we asked participants to select the regulations that would apply to each app, providing both full names and abbreviations of all regulation options. Most participants (70.5%) correctly chose GDPR, while the American privacy regulation CCPA was not chosen as often (26%), although the app descriptions explicitly mentioned that the apps will be published in both European and American markets. Moreover, specialized American regulations—Children's Online Privacy Protection Act (COPPA) [28] and Health Insurance Portability and Accountability Act (HIPAA) [49]—were chosen by 22.8% and 9.9%, respectively, although the described apps were not directed at children and did not collect health-related information.

It is possible that the participants, most of which are from Europe, are more familiar with the European regulations than the American ones, however, we did not find a significant difference between the answers about applicable regulations between the European and North American residents (Mann-Whitney test: $U = 98708.0, p = 0.174$). Finally, 22.8% of participants did not know what regulations apply to the apps, and 2.9% thought that none of them apply. These results show that developers may not be familiar with privacy regulations outside their home country and may not know which regulations are applicable to their apps. It also echos the findings of interviews with developers that they rarely know about privacy guidelines and required measures for privacy [14].

Opinions About User Consent. In the exit survey, we asked participants how they would ask for user consent, assuming they had decided to use personalized ads (Table 5 in the Appendix). The majority (32%) selected the consent form provided by our imaginary Acme ad network. Others preferred to rely on the consent forms provided by leading tech companies (22.5%), such as Facebook or Google, or not-for-profit organizations (10.7%), such as Mozilla or Electronic Frontier Foundation, or use their own consent forms (17.7%). Only 9.75% said they will not ask for user consent at all, assuming that ad network or someone else in the team will take care of it, or because they find the process difficult, unfamiliar, unimportant, or simply not required. Finally, 6% said they would consult the specialized companies providing compliance services.

Information Source	**#Participants**
Reuse available materials	21 (29.6%)
From other companies and not-for-profits	17 (23.9%)
Ready-to-use templates	4 (5.6%)
Guidelines	14 (19.7%)
Legal policies (e.g., GDPR)	10 (14.1%)
UX guidelines	4 (5.6%)
Online search	9 (12.7%)
Legal teams	7 (9.9%)
Relying on own knowledge	6 (8.5%)
Don't know	6 (8.5%)
Unclear responses	12 (16.9%)

Table 4: Constructed themes around participant's information sources for building their consent forms ($N = 71$).

We asked the 71 participants, who indicated they would use their own consent form, what *information sources* they would use to build it (Table 4). After constructing themes around open-ended responses using affinity diagrams, we found that almost a third (29.6%) of participants would still fall back on the existing consent forms built by other teams, apps, companies, non-for-profit organizations, or ready-to-use templates, when building their own forms. Another 19.7% would use general guidelines, such as regulatory policies and recommendations; four participants mentioned using user experience guidelines and best practices when building consent forms: "Existing UX research on consent forms and how to maximize consent with storytelling" (P224).

Other participants said they would search for information about consent forms on the Internet (12.7%), rely on the legal teams or lawyers (9.9%), and their own knowledge or "common sense" (8.5%). However, what constitutes "common sense" for the developer may not necessarily represent what is "common sense" for users. For instance, P277 said that they would tell users that their app uses ads, but would refrain from disclosing that those ads are based on personal information about them: "I'd be upfront about including ads but not state that they dig into people's history" (P277). Finally, 8.5% said they do not know what information they would rely on when building consent forms.

5 Discussion and Future Work

Prior work suggests the importance of improving usability of *security*-related interfaces for developers, for example, through security APIs [43], security notifications [105], and providing secure code examples [69, 70, 71]. Our study highlights the importance of *privacy* interfaces as well by looking at the impact of choice framing on developers' decisions

about user privacy while interacting with ad networks. We hypothesize that the low rate of GDPR-compliant consent forms on websites [37, 67, 112] and the abundance of non-compliant Android apps [60, 89, 96, 118] may partially be caused by developers' low awareness about or consideration of consequences of their decisions on user privacy. We find that incorporating nudges in the design of developers' tools may assist developers in making decisions that consider user privacy in their software development processes.

5.1 Provide Information About Privacy Implications of Ad Personalization

The choice framing that described data processing as being restricted to contextual information instead of past behaviors produced positive but weaker effects compared to the explicit use of privacy labels (11.06 vs. 3.45 times increase in the likelihood to choose non-personalized ads). We believe that this is because in the former case participants had to evaluate themselves the implications of using contextual vs. behavioral targeting on user privacy, while labels that clearly indicated the positive and negative privacy consequences simplified this task. We hypothesize that developers may not fully understand the differences between contextual and behavioral targeting and associated privacy implications; future work is called to explore this hypothesis.

Thus, we recommend ad networks to include information to help developers evaluate privacy implications of their decisions in a transparent, concise, and direct way, by including clear privacy labels to the choices about the ad types. Including these options in the documentation and quick start guides as part of developers' workflow for ads integration may also assist developers in considering user privacy as part of their app development procedure. Additional information on users' concerns about behavioral targeting (e.g., discomforting [63, 117], discriminating [86], and intrusive [83, 88]) might facilitate developers' assessment of privacy implications or support the claims about their relative privacy invasiveness; future work is needed to study how to effectively integrate this information without making the choice text options longer, and whether the manipulation is effective in nudging developers' choices in a less controlled setting.

5.2 Improve the Effectiveness of User-Facing Privacy Descriptions

Prior work recommends emphasizing privacy features in the app stores [59], for instance, the recent inclusion of "Privacy Details" in the Apple App Store aimed at explaining apps' privacy practices before users download them [12]. However, our experiment did not find evidence that adding user-facing descriptions (with our choice framing) of app's ad targeting practices would nudge participants to integrate less invasive non-personalized ads. Participants' open-ended comments suggest a potential explanation: most participants do not expect personalized ads to reduce their app's user base; they also believe that personalized ads are more relevant and less annoying to the users. In other words, some participants believed that telling users that an app shows ads tailored to their personal information will not discourage users from downloading it, and indeed, may even attract users who prefer ads relevant and customized to their interests. However, prior work shows that some users do not like behaviorally targeted ads, find them invasive and creepy, and try to avoid or block such ads [5, 10, 75, 97, 99, 110].

Future work is called for to explore more efficient ways to nudge developers to consider privacy implications of their in-app ad choices. For instance, studying how to best provide evidence to developers about user opinions around ads, privacy preferences, and the impact of app-store presented information, would all help better inform developers' choices. Moreover, future work may test and improve the effectiveness of the existing ways to increase transparency and developers' responsibility to users' regarding their privacy, such as adding "Privacy Details" in the Apple App Store [12], potentially from a privacy nutrition labels perspective [53].

5.3 Reconcile Contradicting Beliefs

As we explained in Section 4.2, the app category did not impact the decisions between the personalized and non-personalized ads, and the number of participants in each group differed only slightly. The analysis of category-related reasons (Section 4.3) provides a potential explanation why we might have not observed a difference. Specifically, it revealed the contradicting beliefs about the same app category that lead to different ad type choices, potentially canceling out the effects of app category. For example, while some participants preferred non-personalized ads for financial apps to avoid raising privacy and trust concerns among users, others preferred to maximize profit from showing the personalized ads to this affluent user group, particularly valued by the advertisers. In the gaming context, because presumably the app does not collect sensitive information, some chose personalized ads as they believed it would not raise privacy concerns, others chose non-personalized ads as it would not be possible to customize ads due to the lack of personal information.

Similar contradictions are revealed in the experimental conditions. When we emphasized privacy implications, the majority of participants chose more privacy-friendly non-personalized ads. When we emphasized the implications on app's revenue, the majority chose revenue-maximizing personalized ads. However, when faced with an explicit choice between user privacy and app's revenue, the choices between two types of ads split almost equally, with a small preference for non-personalized ads. This finding suggests the balance between the contradicting values is fragile and can be easily manipulated. Similar to users' privacy decisions being

context-dependent [2, 4, 80], developers' decisions may also be driven by contextual factors. As some of our participants clarified in the open-ended responses, this choice may change depending on the associated impact on revenue or user privacy. For instance, if the promised increase in revenue is high enough, developers may choose it over user privacy; if they believe that the data collected by the app or context of the app in general is particularly sensitive to raise user concerns, they may be more prone to choose user privacy over profit.

Developers may integrate ad networks primarily because they see it as the only feasible way to monetize the app [68]. The current choice framing in the ad networks also favors the revenue and uses a language that nudges developers into choosing the personalized ads [102]. However, there are also hidden costs of mobile ads that many developers do not consider in weighing the trade-offs, such as frequent updating of ad-related code, and increased consumption of energy and network data on users' phone and subsequent decrease in app's use [44]. Future work could suggest ways to provide transparency about such trade-offs by looking at proposed frameworks for improving the equilibrium between the revenue and user privacy in smartphones by adjusting the level of privacy protection in response to ad-generated revenue [57].

Our results also inform regulators that slight changes in ad networks' interface design for developers may affect the fragile balance between the contradicting values of personalized ads and significantly affect developers' choices to benefit platform's interests in profit maximization. We recommend regulators build clear technical recommendations for providing choices to users, and to enforce that ad networks and other platforms use the mandated framing to promote users' welfare, and avoid effects driven by platforms' sole interests. Future work could provide inputs to the regulators by studying the usability of developer-facing interfaces (e.g., the privacy dashboard on Google AdMob), to inform the design of such interfaces and to provide suggestions to regulators on how to minimize the use of dark patterns in these interfaces.

5.4 Increase Developers' and Users' Control Over Data and Transparency

Many participants said that they do not have full control over ad networks' data collection and processing for ad personalization, and that users have even less control over it. We recommend ad networks, and app stores in particular, to increase the transparency about data practices, accountability to users, and developers' and users' control over data. For instance, Google Play's privacy nudges for permissions has shown success in reducing the number of permissions that developers request [85]. This model might be used to make information about third-party libraries such as ad networks more specific. We suggest app stores to scan for ad libraries and inform developers about their privacy implications during the automatic reviews of the apps (as they currently do for other purposes such as displaying third-party apps [13]).

Some of our participants said that they prefer to let users decide what types of ads they want to see (personalized or non-personalized). However, this line of thought is not completely fair to the users in the environment of information asymmetry, where users are poorly informed about the data practices of apps and ad networks, and personal data flows are not transparent to the users [9, 15, 22, 78]. Thus, providing means for users to see what ad networks are being used in apps when installing a new app [29], what types of ads do the apps serve, and what personal information is used to customize them, as well as other improvement in user interfaces described in Section 5.2, might be effective. Prior results from user research may also help build usable privacy interfaces for developers and increase transparency and control. For instance, several elements of the labels such as data collection, purpose, and data sharing [34, 53] might be reused to inform developers about an ad network's data collection. Other proposed interfaces that visually represent permissions, purposes, data leaks [61, 114], data flows, the effects of removing and adding libraries [113], and integrating privacy checks into programming interfaces [58] might further inform developers about the privacy consequences of their choices. Not-for-profit organizations could build open-source services and easy-to-integrate privacy consent mechanisms to facilitate consent integration, and offer alternatives to for-profit large companies consent forms. Future work could also evaluate the effectiveness of various types of information sources on developers' success in building compliant and user-friendly consent forms (Table 4).

6 Conclusion

We present the results of an online experiment with 400 participants with mobile app development experience on their decisions regarding configuring ads for hypothetical apps. We find that the choice framing in ad networks significantly impacts developers' choices and subsequently privacy of millions of users. Thus, more control and transparency should be provided to developers and users in choosing the type of ads and data collection practices. Moreover, some of our participants incorrectly identified what privacy regulations would apply to the apps, and many said they rely on ad networks and examples from tech companies, when building user consent forms. This means that those companies are not only responsible to their own users, but also set example for other smaller companies and independent developers, further illustrating the large impact of ad network platform's design and choice framing on data practices in app development. Our results have implications for ad networks, app stores, and regulators by giving them grounds for promoting user privacy by improving the usability of developer-facing interfaces to empower developers in making informed decisions for their users.

Acknowledgments

We thank the anonymous reviewers whose comments helped improve the paper greatly. This work was sponsored in part by Microsoft Research through its PhD Scholarship Program, a Google Research Award, and the National Security Agency's Science of Security program. Opinions, findings, and conclusions are those of the authors and do not necessarily reflect the views of the funders.

References

[1] Yasemin Acar, Christian Stransky, Dominik Wermke, Michelle L. Mazurek, and Sascha Fahl. "Security Developer Studies with GitHub Users: Exploring a Convenience Sample". In: *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*. Santa Clara, CA: USENIX Association, July 2017, pp. 81–95. URL: https://www.usenix.org/conference/soups2017/technical-sessions/presentation/acar.

[2] Mark Ackerman, Trevor Darrell, and Daniel J Weitzner. "Privacy in context". In: *Human–Computer Interaction* 16.2-4 (2001), pp. 167–176. DOI: 10.1207/S15327051HCI16234_03.

[3] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. "Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online". In: *ACM Computing Surveys* 50.3 (Aug. 2017). DOI: 10.1145/3054926.

[4] Alessandro Acquisti, Leslie K John, and George Loewenstein. "What is privacy worth?" In: *The Journal of Legal Studies* 42.2 (2013), pp. 249–274. DOI: 10.1086/671754.

[5] *Ad-Blocking: A deep-dive into ad-blocking trends*. Tech. rep. GlobalWebIndex, 2018. URL: https://www.globalwebindex.com/hubfs/Downloads/Ad-Blocking-trends-report.pdf (visited on 02/2021).

[6] Md Ahasanuzzaman, Safwat Hassan, Cor-Paul Bezemer, and Ahmed E. Hassan. "A longitudinal study of popular ad libraries in the Google Play Store". en. In: *Empirical Software Engineering* 25.1 (Jan. 2020), pp. 824–858. DOI: 10.1007/s10664-019-09766-x.

[7] Md Ahasanuzzaman, Safwat Hassan, and Ahmed E. Hassan. "Studying Ad Library Integration Strategies of Top Free-to-Download Apps". In: *IEEE Transactions on Software Engineering* PP (Mar. 2020), pp. 1–1. DOI: 10.1109/TSE.2020.2983399.

[8] Icek Ajzen. "From Intentions to Actions: A Theory of Planned Behavior". In: *Action Control: From Cognition to Behavior*. Ed. by Julius Kuhl and Jürgen Beckmann. Berlin, Heidelberg: Springer Berlin Heidelberg, 1985, pp. 11–39. DOI: 10.1007/978-3-642-69746-3_2.

[9] Urs-Vito Albrecht. "Transparency of health-apps for trust and decision making." eng. In: *Journal of medical Internet research* 15.12 (Dec. 2013). ISSN: 1438-8871 1439-4456. DOI: 10.2196/jmir.2981.

[10] Mimi An. *Why People Block Ads (And What It Means for Marketers and Advertisers)*. HubSpot. 2020. URL: https://blog.hubspot.com/marketing/why-people-block-ads-and-what-it-means-for-marketers-and-advertisers (visited on 02/2021).

[11] *Android Ad Network statistics and market share*. AppBrain. 2020. URL: https://www.appbrain.com/stats/libraries/ad-networks (visited on 09/2020).

[12] *App privacy details on the App Store*. Apple. 2021. URL: https://developer.apple.com/app-store/app-privacy-details/ (visited on 02/2021).

[13] *App Store Review Guidelines*. Apple. 2021. URL: https://developer.apple.com/app-store/review/guidelines/#unacceptable (visited on 02/2021).

[14] Rebecca Balebako and Lorrie Cranor. "Improving App Privacy: Nudging App Developers to Protect User Privacy". In: *IEEE Security Privacy* 12.4 (2014), pp. 55–58. DOI: 10.1109/MSP.2014.70.

[15] Rebecca Balebako, Jaeyeon Jung, Wei Lu, Lorrie Faith Cranor, and Carolyn Nguyen. ""Little Brothers Watching You": Raising Awareness of Data Leaks on Smartphones". In: *Proceedings of the Ninth Symposium on Usable Privacy and Security*. SOUPS '13. Newcastle, United Kingdom: ACM, 2013. DOI: 10.1145/2501604.2501616.

[16] Rebecca Balebako, Pedro G Leon, Hazim Almuhimedi, Patrick Gage Kelley, Jonathan Mugan, Alessandro Acquisti, Lorrie Cranor, and Norman Sadeh-Koniecpol. "Nudging users towards privacy on mobile devices". In: *CHI 2011 Workshop on Persuasion, Nudge, Influence and Coercion*. Carnegie Mellon University, 2011. DOI: 10.1184/R1/13028258.v1.

[17] Rebecca Balebako, Abigail Marsh, Jialiu Lin, Jason I Hong, and Lorrie Cranor. "The privacy and security behaviors of smartphone app developers". In: *Workshop on Usable Security (USEC'14)*. Internet Society, 2014. DOI: 10.14722/usec.2014.23006.

[18] Rebecca Balebako, Florian Schaub, Idris Adjerid, Alessandro Acquisti, and Lorrie Cranor. "The Impact of Timing on the Salience of Smartphone App Privacy Notices". In: *Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices*. SPSM '15. Denver, Colorado, USA: ACM, 2015, pp. 63–74. DOI: 10.1145/2808117.2808119.

[19] Kenneth A Bamberger, Serge Egelman, Catherine Han, Amit Elazari Bar On, and Irwin Reyes. "Can You Pay For Privacy? Consumer Expectations and the Behavior of Free and Paid Apps". In: *Berkeley Technology Law Journal* 35 (2020). DOI: 10.15779/Z38XP6V40J.

[20] Hyejin Bang and Bartosz W. Wojdynski. "Tracking users' visual attention and responses to personalized advertising based on task cognitive demand". In: *Computers in Human Behavior* 55 (2016), pp. 867–876. DOI: 10.1016/j.chb.2015.10.025.

[21] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software, Articles* 67.1 (2015), pp. 1–48. DOI: 10.18637/jss.v067.i01.

[22] Jan Hendrik Betzing, Matthias Tietz, Jan vom Brocke, and Jörg Becker. "The impact of transparency on mobile privacy decision making". In: *Electronic Markets* 30.3 (Sept. 2020), pp. 607–625. ISSN: 1422-8890. DOI: 10.1007/s12525-019-00332-3.

[23] Christoph Bösch, Benjamin Erb, Frank Kargl, Henning Kopp, and Stefan Pfattheicher. "Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns". In: *Proceedings on Privacy Enhancing Technologies* 2016.4 (2016), pp. 237–254. DOI: 10.1515/popets-2016-0038.

[24] Virginia Braun and Victoria Clarke. "Using thematic analysis in psychology". In: *Qualitative Research in Psychology* 3.2 (2006), pp. 77–101. DOI: 10.1191/1478088706qp063oa.

[25] Alex Braunstein, Laura Granka, and Jessica Staddon. "Indirect Content Privacy Surveys: Measuring Privacy without Asking about It". In: *Proceedings of the Seventh Symposium on Usable Privacy and Security*. SOUPS '11. Pittsburgh, Pennsylvania: ACM, 2011. DOI: 10.1145/2078827.2078847.

[26] *California Consumer Privacy Act (CCPA)*. State of California Department of Justice. 2018. URL: https://oag.ca.gov/privacy/ccpa (visited on 09/2020).

[27] Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. "23 Ways to Nudge: A Review of Technology-Mediated Nudging in Human-Computer Interaction". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland UK: ACM, 2019, pp. 1–15. DOI: 10.1145/3290605.3300733.

[28] *Children's Online Privacy Protection Rule (COPPA)*. Federal Trade Commission. 1998. URL: https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule (visited on 09/2020).

[29] Saksham Chitkara, Nishad Gothoskar, Suhas Harish, Jason I. Hong, and Yuvraj Agarwal. "Does This App Really Need My Location? Context-Aware Privacy Management for Smartphones". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.3 (Sept. 2017). DOI: 10.1145/3132029.

[30] Joseph Cox. *How the U.S. Military Buys Location Data from Ordinary Apps*. VICE. 2020. URL: https://www.vice.com/en/article/jgqm5x/us-military-location-data-xmode-locate-x (visited on 02/2021).

[31] *Digital Advertising Alliance (DAA)*. Digital Advertising Alliance (DAA). 2021. URL: https://digitaladvertisingalliance.org (visited on 02/2021).

[32] Serge Egelman, Marian Harbach, and Eyal Peer. "Behavior Ever Follows Intention? A Validation of the Security Behavior Intentions Scale (SeBIS)". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM, 2016, pp. 5257–5261. DOI: 10.1145/2858036.2858265.

[33] Anirudh Ekambaranathan, Jun Zhao, and Max Van Kleek. "Understanding Value and Design Choices Made by Android Family App Developers". In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI EA '20. Honolulu, HI, USA: ACM, 2020, pp. 1–10. DOI: 10.1145/3334480.3383064.

[34] Pardis Emami-Naeini, Yuvraj Agarwal, Lorrie Faith Cranor, and Hanan Hibshi. "Ask the Experts: What Should Be on an IoT Privacy and Security Label?" In: *2020 IEEE Symposium on Security and Privacy (SP)*.

2020, pp. 447–464. DOI: 10.1109/SP40000.2020.00043.

[35] *Explore our participant pool demographics*. Prolific. 2021. URL: https://www.prolific.co/demographics/ (visited on 02/2021).

[36] Robert J Fisher. "Social desirability bias and the validity of indirect questioning". In: *Journal of consumer research* 20.2 (1993), pp. 303–315. DOI: 10.1086/209351.

[37] Imane Fouad, Cristiana Santos, Feras Al Kassar, Nataliia Bielova, and Stefano Calzavara. "On Compliance of Cookie Purposes with the Purpose Specification Principle". In: *IWPE 2020 - International Workshop on Privacy Engineering*. Genova, Italy, Sept. 2020, pp. 1–8. URL: https://hal.inria.fr/hal-02567022.

[38] Andrew Gelman, Masanao Yajima Yu-Sung Su, Jennifer Hill, Maria Grazia Pittau, Jouni Kerman, Tian Zheng, and Vincent Dorie. *arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*. 2020. URL: https://cran.r-project.org/package=arm (visited on 02/2021).

[39] *General Data Protection Regulation (GDPR)*. The European parliament and the council of the European union. 2018. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679 (visited on 09/2020).

[40] Phillipa Gill, Vijay Erramilli, Augustin Chaintreau, Balachander Krishnamurthy, Konstantina Papagiannaki, and Pablo Rodriguez. "Follow the Money: Understanding Economics of Online Aggregation and Advertising". In: *Proceedings of the 2013 Conference on Internet Measurement Conference*. IMC '13. Barcelona, Spain: ACM, 2013, pp. 141–148. DOI: 10.1145/2504730.2504768.

[41] Avi Goldfarb. "What is Different About Online Advertising?" In: *Review of Industrial Organization* 44.2 (Mar. 2014), pp. 115–129. DOI: 10.1007/s11151-013-9399-3.

[42] Colin M. Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L. Toombs. "The Dark (Patterns) Side of UX Design". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. Montreal QC, Canada: ACM, 2018, pp. 1–14. DOI: 10.1145/3173574.3174108.

[43] Matthew Green and Matthew Smith. "Developers Are Not the Enemy!: The Need for Usable Security APIs". In: *IEEE Security and Privacy* 14.5 (Sept. 2016), pp. 40–46. DOI: 10.1109/MSP.2016.111.

[44] Jiaping Gui, Stuart Mcilroy, Meiyappan Nagappan, and William G. J. Halfond. "Truth in Advertising: The Hidden Cost of Mobile Ads for Software Developers". In: *Proceedings of the 37th International Conference on Software Engineering - Volume 1*. ICSE '15. Florence, Italy: IEEE Press, 2015, pp. 100–110. DOI: 10.1109/ICSE.2015.32.

[45] Hana Habib, Sarah Pearman, Jiamin Wang, Yixin Zou, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. ""It's a Scavenger Hunt": Usability of Websites' Opt-Out and Data Deletion Choices". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: ACM, 2020, pp. 1–12. DOI: 10.1145/3313831.3376511.

[46] Hana Habib, Yixin Zou, Aditi Jannu, Neha Sridhar, Chelse Swoopes, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. "An Empirical Analysis of Data Deletion and Opt-Out Choices on 150 Websites". In: *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. Santa Clara, CA: USENIX Association, Aug. 2019. URL: https://www.usenix.org/conference/soups2019/presentation/habib.

[47] Catherine Han, Irwin Reyes, Álvaro Feal, Joel Reardon, Primal Wijesekera, Amit Elazari, Kenneth A Bamberger, and Serge Egelman. "The Price is (Not) Right: Comparing Privacy in Free and Paid Apps". In: *Privacy Enhancing Technologies Symposium*. 2020, p. 21. DOI: 10.2478/popets-2020-0050.

[48] Boyuan He, Haitao Xu, Ling Jin, Guanyu Guo, Yan Chen, and Guangyao Weng. "An Investigation into Android In-App Ad Practice: Implications for App Developers". In: *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications*. 2018, pp. 2465–2473. DOI: 10.1109/INFOCOM.2018.8486010.

[49] *Health Insurance Portability and Accountability Act of 1996 (HIPAA)*. U.S. Department of Health & Human Services. 1996. URL: https://www.cdc.gov/phlp/publications/topic/hipaa.html (visited on 02/2021).

[50] Maximilian Hils, Daniel W. Woods, and Rainer Böhme. "Measuring the Emergence of Consent Management on the Web". In: *Proceedings of the ACM Internet Measurement Conference*. IMC '20. Virtual Event, USA: ACM, 2020, pp. 317–332. DOI: 10.1145/3419394.3423647.

[51] Ling Jin, Boyuan He, Guangyao Weng, Haitao Xu, Yan Chen, and Guanyu Guo. "MAdLens: Investigating into Android In-App Ad Practice at API Granularity". In: *IEEE Transactions on Mobile Computing*

PP.PP (2019), pp. 1–1. DOI: 10.1109/TMC.2019.2953609.

[52] Eric J Johnson, Steven Bellman, and Gerald L Lohse. "Defaults, Framing and Privacy: Why Opting In-Opting Out[1]". In: *Marketing letters* 13.1 (Feb. 2002), pp. 5–15. DOI: 10.1023/A:1015044207315.

[53] Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. "A "Nutrition Label" for Privacy". In: *Proceedings of the 5th Symposium on Usable Privacy and Security*. SOUPS '09. Mountain View, California, USA: ACM, 2009. DOI: 10.1145/1572532.1572538.

[54] Spyros Kokolakis. "Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon". In: *Computers & Security* 64 (2017), pp. 122–134. DOI: 10.1016/j.cose.2015.07.002.

[55] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. "Chapter 8 - Interviews and focus groups". In: *Research Methods in Human Computer Interaction*. Ed. by Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. Second Edition. Boston: Morgan Kaufmann, 2017, pp. 187–228. DOI: 10.1016/B978-0-12-805390-4.00008-X.

[56] Pedro Leon, Blase Ur, Richard Shay, Yang Wang, Rebecca Balebako, and Lorrie Cranor. "Why Johnny Can't Opt out: A Usability Evaluation of Tools to Limit Online Behavioral Advertising". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '12. Austin, Texas, USA: ACM, 2012, pp. 589–598. DOI: 10.1145/2207676.2207759.

[57] Ilias Leontiadis, Christos Efstratiou, Marco Picone, and Cecilia Mascolo. "Don't Kill My Ads! Balancing Privacy in an Ad-Supported Mobile Application Market". In: *Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications*. HotMobile '12. San Diego, California: ACM, 2012. DOI: 10.1145/2162081.2162084.

[58] Tianshi Li, Yuvraj Agarwal, and Jason I. Hong. "Coconut: An IDE Plugin for Developing Privacy-Friendly Apps". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.4 (Dec. 2018). DOI: 10.1145/3287056.

[59] Tianshi Li, Elizabeth Louie, Laura Dabbish, and Jason I. Hong. "How Developers Talk About Personal Data and What It Means for User Privacy: A Case Study of a Developer Forum on Reddit". In: *Proc. ACM Hum.-Comput. Interact.* 4.CSCW (Jan. 2021). DOI: 10.1145/3432919.

[60] Ilaria Liccardi, Monica Bulger, Hal Abelson, Daniel Weitzner, and Wendy Mackay. "Can apps play by the COPPA Rules?" In: *2014 Twelfth Annual International Conference on Privacy, Security and Trust*. 2014, pp. 1–9. DOI: 10.1109/PST.2014.6890917.

[61] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi, Shikun (Aerin) Zhang, Norman Sadeh, Yuvraj Agarwal, and Alessandro Acquisti. "Follow My Recommendations: A Personalized Privacy Assistant for Mobile App Permissions". In: *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. Denver, CO: USENIX Association, June 2016, pp. 27–41. URL: https://www.usenix.org/conference/soups2016/technical-sessions/presentation/liu.

[62] Daniel Lüdecke, Dominique Makowski, Philip Waggoner, and Indrajeet Patil. *performance: Assessment of Regression Models Performance*. R package. 2020. DOI: 10.5281/zenodo.3952174.

[63] Miguel Malheiros, Charlene Jennett, Snehalee Patel, Sacha Brostoff, and Martina Angela Sasse. "Too Close for Comfort: A Study of the Effectiveness and Acceptability of Rich-Media Personalized Advertising". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '12. Austin, Texas, USA: ACM, 2012, pp. 579–588. DOI: 10.1145/2207676.2207758.

[64] Miriam Marciel, J. G. Cabañas, Y. Kassa, R. Gonzalez, and M. Ahmed. "The Value of Online Users: Empirical Evaluation of the Price of Personalized Ads". In: *2016 11th International Conference on Availability, Reliability and Security (ARES)*. 2016, pp. 694–700. DOI: 10.1109/ARES.2016.89.

[65] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. "Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites". In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (Nov. 2019). DOI: 10.1145/3359183.

[66] Arunesh Mathur, Jessica Vitak, Arvind Narayanan, and Marshini Chetty. "Characterizing the Use of Browser-Based Blocking Extensions To Prevent Online Tracking". In: *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*. Baltimore, MD: USENIX Association, Aug. 2018, pp. 103–116. URL: https://www.usenix.org/conference/soups2018/presentation/mathur.

[67] Celestin Matte, Nataliia Bielova, and Cristiana Santos. "Do Cookie Banners Respect my Choice? : Measuring Legal Compliance of Banners from IAB Europe's Transparency and Consent Framework". In:

2020 IEEE Symposium on Security and Privacy (SP). May 2020, pp. 791–809. DOI: 10.1109/SP40000.2020.00076.

[68] Abraham H. Mhaidli, Yixin Zou, and Florian Schaub. ""We Can't Live Without Them!" App Developers' Adoption of Ad Networks and Their Considerations of Consumer Risks". In: *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. Santa Clara, CA: USENIX Association, Aug. 2019. URL: https://www.usenix.org/conference/soups2019/presentation/mhaidli.

[69] Kai Mindermann, Philipp Keck, and Stefan Wagner. "How Usable Are Rust Cryptography APIs?" In: *2018 IEEE International Conference on Software Quality, Reliability and Security (QRS)*. IEEE, July 2018, pp. 143–154. DOI: 10.1109/qrs.2018.00028.

[70] Kai Mindermann and Stefan Wagner. "Fluid Intelligence Doesn't Matter! Effects of Code Examples on the Usability of Crypto APIs". In: *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: Companion Proceedings*. ICSE '20. Seoul, South Korea: ACM, 2020, pp. 306–307. DOI: 10.1145/3377812.3390892.

[71] Kai Mindermann and Stefan Wagner. "Usability and Security Effects of Code Examples on Crypto APIs". In: *2018 16th Annual Conference on Privacy, Security and Trust (PST)*. IEEE, Aug. 2018, pp. 1–2. DOI: 10.1109/PST.2018.8514203.

[72] *Miro | Online Whiteboard for Visual Collaboration*. Miro. 2021. URL: https://miro.com/ (visited on 02/2021).

[73] *Mixed Effects Logistic Regression*. UCLA: Statistical Consulting Group. 2020. URL: https://stats.idre.ucla.edu/stata/dae/mixed-effects-logistic-regression/ (visited on 02/2021).

[74] *MoPub Integration Suite*. Twitter MoPub. 2021. URL: https://developers.mopub.com/publishers/integrate/ (visited on 02/2021).

[75] Lymari Morales. *U.S. Internet Users Ready to Limit Online Tracking for Ads*. Gallup Polls. 2010. URL: https://news.gallup.com/poll/145337/internet-users-ready-limit-online-tracking-ads.aspx (visited on 02/2021).

[76] *Most popular Apple App Store categories in August 2020, by share of available apps*. Statista. 2020. URL: https://www.statista.com/statistics/270291/popular-categories-in-the-app-store/ (visited on 02/2021).

[77] *Most popular Google Play app categories as of 3rd quarter 2020, by share of available apps*. Statista. 2020. URL: https://www.statista.com/statistics/279286/google-play-android-app-categories/ (visited on 02/2021).

[78] Patrick Murmann. "Usable Transparency for Enhancing Privacy in Mobile Health Apps". In: *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. MobileHCI '18. Barcelona, Spain: ACM, 2018, pp. 440–442. DOI: 10.1145/3236112.3236184.

[79] Arvind Narayanan, Arunesh Mathur, Marshini Chetty, and Mihir Kshirsagar. "Dark Patterns: Past, Present, and Future". In: *Queue* 18.2 (Apr. 2020), pp. 67–92. DOI: 10.1145/3400899.3400901.

[80] Helen Nissenbaum. "Privacy as contextual integrity". In: *Wash. L. Rev.* 79 (2004), p. 119.

[81] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. "Dark Patterns after the GDPR: Scraping Consent Pop-Ups and Demonstrating Their Influence". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: ACM, 2020, pp. 1–13. DOI: 10.1145/3313831.3376321.

[82] *Number of software developers worldwide in 2018, 2019, 2023 and 2024*. Statista. 2020. URL: https://www.statista.com/statistics/627312/worldwide-developer-population/ (visited on 02/2021).

[83] Katie O'Donnell and Henriette Cramer. "People's Perceptions of Personalized Ads". In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15 Companion. Florence, Italy: ACM, 2015, pp. 1293–1298. DOI: 10.1145/2740908.2742003.

[84] *Out of Control - How consumers are exploited by the online advertising industry*. Forbrukerrådet. 2020. URL: https://www.forbrukerradet.no/side/new-study-the-advertising-industry-is-systematically-breaking-the-law (visited on 09/2020).

[85] Sai Teja Peddinti, Igor Bilogrevic, Nina Taft, Martin Pelikan, Úlfar Erlingsson, Pauline Anthonysamy, and Giles Hogben. "Reducing Permission Requests in Mobile Apps". In: *Proceedings of the Internet Measurement Conference*. IMC '19. Amsterdam, Netherlands: ACM, 2019, pp. 259–266. DOI: 10.1145/3355369.3355584.

[86] Angelisa C. Plane, Elissa M. Redmiles, Michelle L. Mazurek, and Michael Carl Tschantz. "Exploring User Perceptions of Discrimination in Online Targeted Advertising". In: *26th USENIX Security Symposium (USENIX Security 17)*. Vancouver, BC: USENIX Association, Aug. 2017, pp. 935–951. URL: https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/plane.

[87] Isaac Prilleltensky. "Psychology and the status quo." In: *American Psychologist* 44.5 (1989), pp. 795–802. DOI: 10.1037/0003-066X.44.5.795.

[88] *Public Attitudes Towards Online Targeting*. Centre for Data Ethics and Innovation. 2020. URL: https://www.gov.uk/government/publications/cdei-review-of-online-targeting (visited on 02/2021).

[89] Irwin Reyes, Primal Wijesekera, Joel Reardon, Amit Elazari Bar On, Abbas Razaghpanah, Narseo Vallina-Rodriguez, and Serge Egelman. ""Won't Somebody Think of the Children?" Examining COPPA Compliance at Scale". In: *Proceedings on Privacy Enhancing Technologies* 2018.3 (2018), pp. 63–83. DOI: 10.1515/popets-2018-0021.

[90] Takahito Sakamoto and Masahiro Matsunaga. "After GDPR, Still Tracking or Not? Understanding Opt-Out States for Online Behavioral Advertising". In: *2019 IEEE Security and Privacy Workshops (SPW)*. 2019, pp. 92–99. DOI: 10.1109/SPW.2019.00027.

[91] William Samuelson and Richard Zeckhauser. "Status quo bias in decision making". In: *Journal of risk and uncertainty* 1.1 (Mar. 1988), pp. 7–59. DOI: 10.1007/BF00055564.

[92] Deborah H Schenk. "Exploiting the Salience Bias in Designing Taxes". In: *Yale J. on Reg.* 28 (2011), p. 253. DOI: 10.2139/ssrn.1661322.

[93] *Share of global smartphone shipments by operating system from 2014 to 2023*. Statista. 2020. URL: https://www.statista.com/statistics/272307/market-share-forecast-for-smartphone-operating-systems/ (visited on 09/2020).

[94] Swapneel Sheth, Gail Kaiser, and Walid Maalej. "Us and Them: A Study of Privacy Requirements Across North America, Asia, and Europe". In: *Proceedings of the 36th International Conference on Software Engineering*. ICSE 2014. Hyderabad, India: ACM, 2014, pp. 859–870. DOI: 10.1145/2568225.2568244.

[95] Katie Shilton and Daniel Greene. "Linking Platforms, Practices, and Developer Ethics: Levers for Privacy Discourse in Mobile Application Development". In: *Journal of Business Ethics* 155.1 (Mar. 2019), pp. 131–146. DOI: 10.1007/s10551-017-3504-8.

[96] Laura Shipp and Jorge Blasco. "How private is your period?: A systematic analysis of menstrual app privacy policies". In: *Proceedings on Privacy Enhancing Technologies* 2020.4 (Oct. 2020), pp. 491–510. DOI: 10.2478/popets-2020-0083.

[97] Ashish Kumar Singh and Vidyasagar Potdar. "Blocking Online Advertising - A State of the Art". In: *Proceedings of the 2009 IEEE International Conference on Industrial Technology*. ICIT '09. USA: IEEE Computer Society, 2009, pp. 1–10. DOI: 10.1109/ICIT.2009.4939739.

[98] *Software developer gender distribution worldwide*. Statista. 2020. URL: https://www.statista.com/statistics/1126823/worldwide-developer-gender/ (visited on 06/2021).

[99] *Special Eurobarometer 431 "Data protection"*. Tech. rep. European Commission, 2015. URL: https://ec.europa.eu/commfrontoffice/publicopinion/archives/ebs/ebs_431_en.pdf (visited on 02/2021).

[100] Ryan Stevens, Clint Gibler, Jon Crussell, Jeremy Erickson, and Hao Chen. "Investigating User Privacy in Android Ad Libraries". In: *Workshop on Mobile Security Technologies (MoST)*. 2012. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.298.7556.

[101] Mohammad Tahaei, Alisa Frik, and Kami Vaniea. "Privacy Champions in Software Teams: Understanding Their Motivations, Strategies, and Challenges". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. New York, NY, USA: ACM, 2021, pp. 1–15. DOI: 10.1145/3411764.3445768.

[102] Mohammad Tahaei and Kami Vaniea. ""Developers Are Responsible": What Ad Networks Tell Developers About Privacy". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems Extended Abstracts*. CHI '21 Extended Abstracts. New York, NY, USA: ACM, 2021, pp. 1–12. DOI: 10.1145/3411763.3451805.

[103] Mohammad Tahaei and Kami Vaniea. "A Survey on Developer-Centred Security". In: *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. June 2019, pp. 129–138. DOI: 10.1109/EuroSPW.2019.00021.

[104] Mohammad Tahaei and Kami Vaniea. "Code-Level Dark Patterns: Exploring Ad Networks' Misleading Code Samples with Negative Consequences for Users". In: *What Can CHI Do About Dark Patterns? Workshop at CHI '21*. 2021, pp. 1–5. URL: http://hdl.handle.net/20.500.11820/ea71877b-4def-4c2c-aa45-e148122b4f36.

[105] Mohammad Tahaei, Kami Vaniea, Beznosov Konstantin, and Maria K. Wolters. "Security Notifications in Static Analysis Tools: Developers' Attitudes, Comprehension, and Ability to Act on Them". In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. CHI '21. New York, NY, USA: ACM, 2021, pp. 1–17. DOI: 10.1145/3411764.3445616.

[106] Mohammad Tahaei, Kami Vaniea, and Naomi Saphra. "Understanding Privacy-Related Questions on Stack Overflow". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. CHI '20. Honolulu, HI, USA: ACM, 2020, pp. 1–14. DOI: 10.1145/3313831.3376768.

[107] *The State of Mobile in 2020*. App Annie. 2020. URL: https://www.appannie.com/en/insights/market-data/state-of-mobile-2020/ (visited on 09/2020).

[108] *The Value of Personalized Ads to a Thriving App Ecosystem*. Facebook. 2020. URL: https://developers.facebook.com/blog/post/2020/06/18/value-of-personalized-ads-thriving-app-ecosystem/ (visited on 02/2021).

[109] Janice Y. Tsai, Serge Egelman, Lorrie Cranor, and Alessandro Acquisti. "The Effect of Online Privacy Information on Purchasing Behavior: An Experimental Study". In: *Info. Sys. Research* 22.2 (June 2011), pp. 254–268. DOI: 10.1287/isre.1090.0260.

[110] Joseph Turow, Jennifer King, Chris Hoofnagle, Amy Bleakley, and Michael Hennessy. "Americans Reject Tailored Advertising and Three Activities That Enable It". In: (Sept. 2009). DOI: 10.2139/ssrn.1478214.

[111] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. "Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising". In: *Proceedings of the Eighth Symposium on Usable Privacy and Security*. SOUPS '12. Washington, D.C.: ACM, 2012. DOI: 10.1145/2335356.2335362.

[112] Christine Utz, Martin Degeling, Sascha Fahl, Florian Schaub, and Thorsten Holz. "(Un)Informed Consent: Studying GDPR Consent Notices in the Field". In: *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. CCS '19. London, United Kingdom: ACM, 2019, pp. 973–990. DOI: 10.1145/3319535.3354212.

[113] Max Van Kleek, Reuben Binns, Jun Zhao, Adam Slack, Sauyon Lee, Dean Ottewell, and Nigel Shadbolt. "X-Ray Refine: Supporting the Exploration and Refinement of Information Exposure Resulting from Smartphone Apps". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. CHI '18. ACM, 2018, pp. 1–13. DOI: 10.1145/3173574.3173967.

[114] Max Van Kleek, Ilaria Liccardi, Reuben Binns, Jun Zhao, Daniel J. Weitzner, and Nigel Shadbolt. "Better the Devil You Know: Exposing the Data Sharing Practices of Smartphone Apps". In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. CHI '17. ACM, 2017, pp. 5208–5220. DOI: 10.1145/3025453.3025556.

[115] Haoyu Wang, Zhe Liu, Yao Guo, Xiangqun Chen, Miao Zhang, Guoai Xu, and Jason Hong. "An Explorative Study of the Mobile App Ecosystem from App Developers' Perspective". In: *Proceedings of the 26th International Conference on World Wide Web*. WWW '17. Perth, Australia: International World Wide Web Conferences Steering Committee, 2017, pp. 163–172. DOI: 10.1145/3038912.3052712.

[116] Ying Wang, Ebru Genc, and Gang Peng. "Aiming the Mobile Targets in a Cross-Cultural Context: Effects of Trust, Privacy Concerns, and Attitude". In: *International Journal of Human–Computer Interaction* 36.3 (2020), pp. 227–238. DOI: 10.1080/10447318.2019.1625571.

[117] Jay (Hyunjae) Yu and Brenda Cude. "'Hello, Mrs. Sarah Jones! We recommend this product!' Consumers' perceptions about personalized advertising: comparisons across advertisements delivered via three different types of media". In: *International Journal of Consumer Studies* 33.4 (2009), pp. 503–514. DOI: 10.1111/j.1470-6431.2009.00784.x.

[118] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N. Cameron Russell, and Norman Sadeh. "MAPS: Scaling Privacy Compliance Analysis to a Million Apps". In: *Proceedings on Privacy Enhancing Technologies* 2019.3 (2019), pp. 66–86. DOI: 10.2478/popets-2019-0037.

Appendices

A Survey Instruments

A.1 Screening Survey

1. Please select the statement that best describes your primary role at your current or most recent job.

   • I'm not employed • Jobs NOT related to computer science, informatics, computer engineering, or related fields • Designing products (e.g. UI designer, interaction designer) • Developing software (e.g. programmer, developer, web developer, software engineer) • Testing software (e.g. tester, quality analyst, automation engineer) • Managing software development (e.g. project manager, IT manager, scrum master) • Privacy and/or security engineering (e.g. security engineer, privacy engineer, penetration tester, ethical hacker, cryptographer) • Others (please specify)

2. How many years of experience do you have in software development? (numbers only)

3. How many years have you worked in mobile app development, specifically? (numbers only)

4. How many mobile apps have you worked on in the last 3 years? (numbers only)

A.2 Main Survey

[After the participant read the participant information sheet and consent form, and agreed to participant in the study.]

1. How many mobile apps have you worked on in the last 3 years? (numbers only)

2. *[Scenario description.]* Imagine that you are a shareholder in a software development company. Together with a small team, you created a [personal finance management/gaming] app. The app will be published in Europe and the United States and is mainly targeted towards adults (above age of 18). To monetise the app, you have decided to use the "Acme" ad network to show ads to your users.

   The Acme ad network offers a step-by-step Assistant – a graphical user interface that provides various configuration choices for integrating ads into your [personal finance management/gaming] app. The Assistant asks the developer several questions and then provides ad network configuration code based on the answers that can be imported directly into an app with minimal additional coding required.

   The following are the 5 questions asked by Acme's Assistant, please answer them as if you were developing the [personal finance management/gaming] app.

   I Which ad formats are you integrating?

   • Banner: A basic ad format that appears at the top & bottom of the device screen. • Interstitial: full-page ads appear at natural breaks & transitions, such as level completion. Supports video content. • Rewarded Video: ads reward users for watching short videos and interacting with playable ads and surveys. Good for monetizing free-to-play users. Supports video content. • Native: customisable ad format that matches the look & feel of your app. Ads appear inline with app content. Supports video content.

   II What level of graphics do you want for your ads?

   • Ads with highest graphics quality. These ads will work best on newer phones with the latest operating systems. • Ads with moderate to low graphics quality. These ads will work on most phones.

   III Which platform are you integrating Acme ad network on?

   • Android • iOS • Unity • Windows Phone

   IV Select the type of ads that you want to show. [Participants were asked to choose between the personalized and non-personalized ads described according to the condition, to which they were randomly assigned. See the text of the options in section 3.1.]

   V Which of the following regulations apply to this app?

   • GDPR (General Data Protection Regulation) • CCPA (California Consumer Privacy Act) • COPPA (Children's Online Privacy Protection Act) • HIPAA (Health Insurance Portability and Accountability Act) • None of the above • I don't know

   VI What was the biggest reason that made you pick the ad type: [chosen ads type]? (Please provide at much as details you can. Your response helps us better understand the reasons behind your choices.) [Open-ended question]

   [Repeat the above questions for the second scenario.]

3. Assume that you decided to use personalized ads in both the gaming and financial management apps described earlier. How do you imagine you would go about asking for user consent for the personalized ads?

   • I'd use my own consent form • I'd use the consent form provided by the Acme ad network • I'd use a third-party consent form provided by a leading tech company (e.g., Facebook, Google, Amazon, Twitter) • I'd use a third-party consent form provided by a not-for-profit organization (e.g., Mozilla, Electronic Frontier Foundation) • I'd use a third-party consent form provided by other companies providing compliance services (e.g. OneTrust) • I won't ask for user consent because I don't think it's required • I won't ask for user consent because I don't think it's important • I won't ask for user consent because someone else in the team should take care of it • I won't ask for user consent because it's hard to do so • I won't ask for user consent because I'm not familiar with the consent process • I won't ask for user consent because the Acme ad network will take care of it • Other (please explain)

4. *[If "I'd use my own consent form" chosen.]* What information sources, if any, would you use to build your own consent form? [Open-ended question]

5. How, if at all, would your app's **revenue** change if you chose personalized ads over non-personalized ads in the **[personal financial management/gaming]** app described earlier? [Participants were asked about both app categories, in randomized order.]

• Decrease by more than 81% • Decrease by 61%-80% • Decrease by 41%-60% • Decrease by 21%-40% • Decrease by 1%-20% • It won't change • Increase by 1%-20% • Increase by 21%-40% • Increase by 41%-60% • Increase by 61%-80% • Increase by more than 81%

6. How, if at all, would the number of **users** of your app change if you chose personalized ads over non-personalized ads in the **[personal financial management/gaming]** app described earlier? [Participants were asked about both app categories, in randomized order.]

   • Decrease by more than 81% • Decrease by 61%-80% • Decrease by 41%-60% • Decrease by 21%-40% • Decrease by 1%-20% • It won't change • Increase by 1%-20% • Increase by 21%-40% • Increase by 41%-60% • Increase by 61%-80% • Increase by more than 81%

7. How much priority do you give to privacy improvement and maintenance tasks in your daily development routines?

   • Not a priority • Low priority • Medium priority • High priority • Essential

8. As a developer, how much control do **you** generally have over the amount of data collected by ad networks?

   • No control at all • Very little control • Moderate control • A lot of control • Full control

9. How much control do **users** generally have over the amount of data collected by ad networks?

   • No control at all • Very little control • Moderate control • A lot of control • Full control

10. What platforms have you previously developed apps for?

    • Android • iOS • Blackberry • Windows Phone

11. How involved have you been in in-app advertising activities? [Options were: Not at all, A little, A moderate amount, A lot, A great deal]

    • Choosing an advertising partner or advertising network for an app. • Configuring the types of in-app ads shown in an app (e.g., where to place ads, what categories of ads to show, etc.) • Integrating the necessary code into an app to enable in-app advertising. • Other (please specify)

12. Regarding mobile apps, have you used or worked with any advertising networks?

    • AdColony • Amazon Mobile Ad Network • Facebook Audience Network • Flurry • Google AdMob • InMobi • Millennial media • Twitter MoPub • Unity Ads • Vungle • Greyfriars Bobby • I have never included any ad networks in my mobile apps

13. *[If "I have never included any ad networks in my mobile apps" chosen.]* What are the primary reasons that you never included any ad networks in your apps? (Please provide at much as details you can. Your response helps us better understand your reasons behind your choices.)

14. What is the revenue model of the apps that you typically develop?

    • Free with In-App Advertising, users cannot pay a fee to remove advertisements • Free with In-App Advertising, users can pay a fee to remove advertisements • Freemium model (app is free, certain features cost user's money) • Paid download • In-App purchases (selling physical or virtual goods through the app) • Subscription (similar to Freemium, except instead of paying for extra features, users must pay for extra content) • My apps are completely free • Cannot remember • Other (please specify)

15. Who decides what revenue model to use in the apps that you develop?

    • Only me • Developer(s) / Programmer(s) • Project manager(s) • CEO and/or other upper-level management • Investor(s) • Other (please specify) • I do not know who was involved in the decision process

16. Who decides what advertisement network to use in the apps that you develop?

    • Only me • Developer(s) / Programmer(s) • Project manager(s) • CEO and/or other upper-level management • Investor(s) • Other (please specify) • I do not know who was involved in the decision process

17. What is your main source of income in software or mobile development?

    • I don't make money from software or mobile development • Salary, not dependent on software/app revenue • Primarily salary and bonuses, partially dependent on software/app revenue • Primarily direct software/app revenue • Other (please specify)

18. What type of employment best describes your most recent app development experience?

    • Full time employee (or contractor equivalent) • Part-time employee (or contractor equivalent) • Freelance/consultant • Furloughed (temporarily laid off) or on leave • Unemployed • Student • Retired • Other (please specify)

19. Please select the statement that best describes your primary roles at your most recent job.

    • I'm not employed • Jobs NOT related to computer science, informatics, computer engineering, or related fields • Designing products (e.g. UI designer, interaction designer) • Developing software (e.g. programmer, developer, web developer, software engineer) • Testing software (e.g. tester, quality analyst, automation engineer) • Managing software development (e.g. project manager, IT manager, scrum master) • Privacy and/or security engineering (e.g. security engineer, privacy engineer, penetration tester, ethical hacker, cryptographer) • Others

20. How many years of experience do you have in software development? (numbers only)

21. How many years have you worked in mobile app development specifically? (numbers only)

22. Where did you mainly learn to program and develop software?

    • Self-taught • High school courses • College or university courses • Online courses • Industry or on-the-job training • Others

23. How many people were employed in the organization for which you worked on the app development most recently?

    • 1-9 employees • 10-99 employees • 100-999 employees • 1,000-9,999 employees • 10,000+ employees

24. How many members were in the team that you have directly worked with most recently? (numbers only)

25. How old are you? (numbers only)

26. In which country do you currently reside? [List of countries]

27. If you can't find your country in the above question options, please enter it here. [Open-ended question]

28. What is your gender?

    • Male • Female • Non-binary • Prefer not to say • Prefer to self describe

29. If you'd like to be included in the raffle, please provide your email address.

30. Do you have comments or anything to say about the survey or study in general? (optional)

B Ads Personalization Options on Google AdMob Developer Dashboard

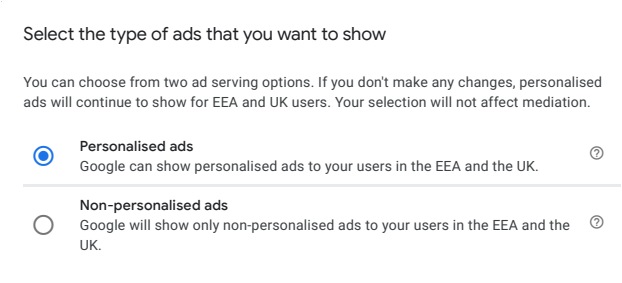


Figure 2: Screenshot from Google AdMob developer dashboard: Blocking controls > Manage EU user consent (Jan'21).

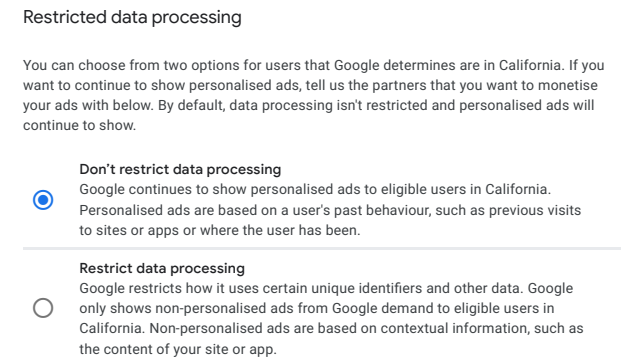


Figure 3: Screenshot of Google AdMob developer dashboard: Blocking controls > Manage CCPA settings (Jan'21).

C Participants' Demographics and Opinions About Ad Networks

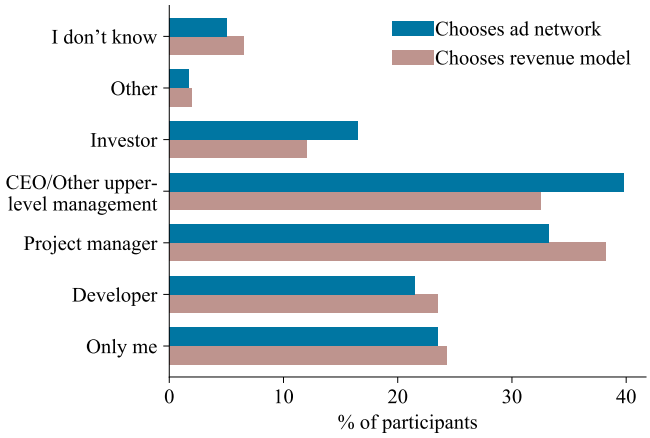


Figure 4: Responses about who decides what revenue model and ad network to use in the apps participants develop.

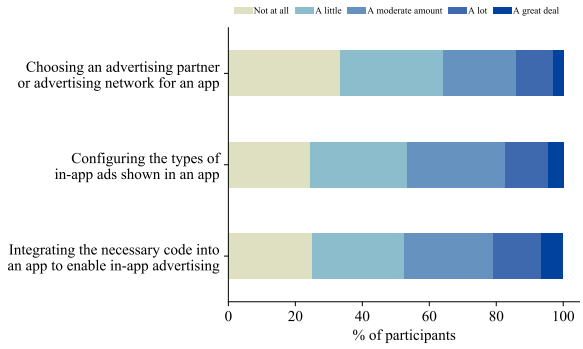


Figure 5: Involvement in in-app advertising activities.

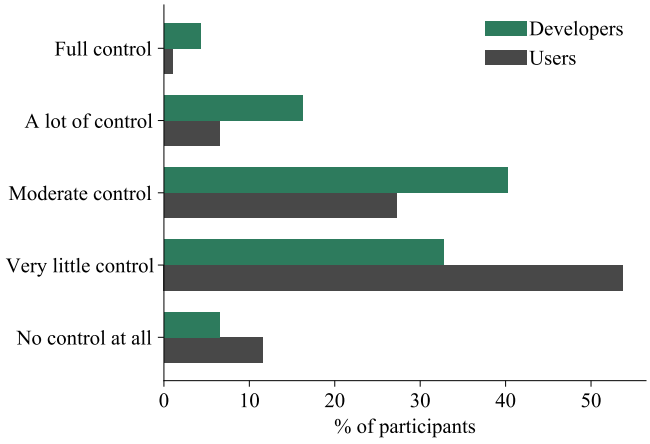


Figure 6: Perceived control over ad networks' data collection.

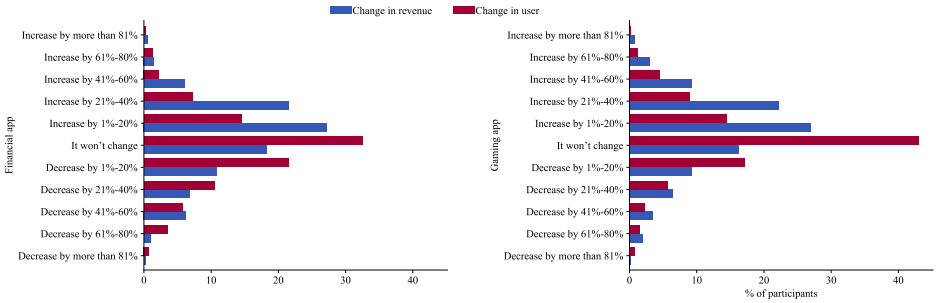


Figure 7: Expected change in app's revenue and N of users if personalized ads are chosen over non-personalized ads.

	#Participants
Age	$\mu = 27.4, \sigma = 8$
Gender	
Male	330 (82%)
Female	58 (14%)
Prefer not to say	11 (3%)
Non-binary	1 (<1%)
Current Continent of Residence	
Europe	265 (66%)
North America	75 (19%)
Asia	24 (6%)
Oceania	15 (4%)
South America	11 (3%)
Africa	7 (2%)
Prefer not to say	3 (1%)
Employment Status	
Full-time	147 (37%)
Student	107 (27%)
Freelance/consultant	75 (19%)
Part-time	54 (14%)
Unemployed	10 (2%)
Temporarily laid off	3 (1%)
Other	2 (<1%)
Retired	2 (<1%)
Number of Employees	
1–9 employees	170 (42%)
10–99 employees	142 (36%)
100–999 employees	49 (12%)
1,000–9,999 employees	21 (5%)
10,000 or more employees	18 (4%)
Team Members	$\mu = 7.3, \sigma = 10.3$
Years of Experience	
In software development	$\mu = 5.1, \sigma = 5.3$
In mobile development	$\mu = 2.7, \sigma = 2.6$
Number of Developed Apps in the Past Three Years	$\mu = 3.5, \sigma = 4.2$
Software-Related Roles ($N = 291$)	
Developing software	186 (64%)
Testing software	37 (13%)
Managing software development	32 (11%)
Designing products	30 (10%)
Privacy & security engineering	5 (2%)
Main Income Source	
Salary, not dependent on software/app revenue	172 (43%)
Salary, partially dependent on software/app revenue	85 (21%)
I don't make money from software/app dev.	80 (20%)
Direct software/app revenue	58 (14%)
Other	5 (1%)

	#Participants
Revenue Models	
Free with In-App Advertising	120 (30.0%)
Completely free	103 (25.8%)
Freemium model	103 (25.8%)
In-App purchases	83 (20.8%)
Free with In-App Advertising	82 (20.5%)
Subscription	54 (13.5%)
Paid download	43 (10.8%)
Other	11 (2.8%)
Can't remember	8 (2.0%)
Which Ad Networks Used in the Past	
Google AdMob	191 (47.8%)
Never included any ad networks in apps	123 (30.8%)
Facebook Audience Network	117 (29.2%)
Unity Ads	81 (20.2%)
Amazon Mobile Ad Network	64 (16.0%)
AdColony	33 (8.2%)
Twitter MoPub	27 (6.8%)
Flurry	15 (3.8%)
InMobi	12 (3.0%)
Other	11 (2.8%)
Vungle	9 (2.2%)
Millennial media	7 (1.8%)
Sources of User Consent Forms	
The Acme ad network's form	128 (32.0%)
Leading tech company's form	90 (22.5%)
My own consent form (see Table 4)	71 (17.8%)
Not-for-profit organization's form	43 (10.8%)
Won't ask for user consent because:	39 (9.75%)
Acme ad network will take care of it	14 (3.5%)
Someone else in the team should do it	14 (3.5%)
Not familiar with the consent process	6 (1.5%)
It's not important	2 (0.5%)
It's hard to do so	2 (0.5%)
It's not required	1 (0.2%)
Companies providing compliance services	24 (6.0%)
Other	5 (1.2%)
Given Priority to Privacy in Development Routines	
High priority	144 (36%)
Medium priority	136 (34%)
Essential	61 (15%)
Low priority	54 (14%)
Not a priority	5 (1%)
Where Learned to Develop Software	
Self-taught	248 (62.0%)
College or university courses	237 (59.2%)
Online courses	170 (42.5%)
Industry or on-the-job training	103 (25.8%)
High school courses	70 (17.5%)
Other	3 (0.8%)

Table 5: Summary of participants' demographics and prior experience with ads ($N = 400$, unless otherwise specified).

Benefits and Drawbacks of Adopting a Secure Programming Language: Rust as a Case Study

Kelsey R. Fulton, Anna Chan, Daniel Votipka[†], Michael Hicks, and Michelle L. Mazurek
University of Maryland [†]*Tufts University*

Abstract

Programming languages such as Rust and Go were developed to combat common and potentially devastating memory-safety-related vulnerabilities. But adoption of new, more secure languages can be fraught and complex. To better understand the benefits and challenges of adopting Rust in particular, we conducted semi-structured interviews with professional, primarily senior software developers who have worked with Rust on their teams or tried to introduce it (n = 16), and we deployed a survey to the Rust development community (n = 178). We asked participants about their personal experiences using Rust, as well as experiences using Rust at their companies. We find a range of positive features, including good tooling and documentation, benefits for the development lifecycle, and improvement of overall secure coding skills, as well as drawbacks including a steep learning curve, limited library support, and concerns about the ability to hire additional Rust developers in the future. Our results have implications for promoting the adoption of Rust specifically and secure programming languages and tools more generally.

1 Introduction

Secure software development is a difficult and important task. Vulnerabilities are still discovered in production code on a regular basis [4, 27, 38], and many of these arise from highly dangerous violations of memory safety, such as use-after-frees, buffer overflows, and out-of-bounds reads/writes [28–32]. Despite their long history and the many attempts aimed at mitigating or blocking their exploitation, such vulnerabilities have remained a consistent, and sometimes worsening, threat [37], with estimates that 60-70% of critical vulnerabilities in Chrome [13], Microsoft products [7] and in other large critical systems [12] owe to memory safety vulnerabilities.

Overwhelmingly, memory safety vulnerabilites occur in C and C++ code—while most popular languages enforce memory safety automatically, C and C++ do not [43, 47]. Relatively recently, Google developed Go [14] and Mozilla developed Rust [33] to be practical but secure alternatives to C and C++; these languages aim to be fast, low-level, *and* type- and memory-safe [34, 40]. Rust and Go have been rising in popularity—IEEE's 2019 Top Programming languages list ranks them 17 and 10, respectively—but C and C++ continue to occupy top spots (3 and 4). We might wonder: What are the factors fueling the rise of these secure languages? Is there a chance they will overtake their insecure counterparts, C and C++, and if so, how?

In this paper, we attempt to answer these questions for Rust, in particular. While Go is extremely popular, Rust's popularity has also risen sharply in the last few years [9, 15, 34, 39, 46]. Rust's "zero-cost abstractions" and its lack of garbage collection make it appropriate for resource-constrained environments, where Go would be less appropriate and C and C++ have traditionally been the only game in town.

We conducted semi-structured interviews with professional, primarily senior developers who have actively worked with Rust on their product teams, and/or attempted to get their companies to adopt Rust (n = 16). We also surveyed participants in Rust development community forums (n = 178). We asked participants about their general programming experience and experiences using and adopting Rust both personally and at a company. We also asked about the benefits and drawbacks of using Rust in both settings. By asking these questions, we aim to understand the challenges that inhibit adoption, the net benefits (if any) that accrue after adoption, and what tactics have been (un)successful in driving adoption and use.

Our survey population likely represents those who view Rust at least somewhat positively, since we would not expect those who tried Rust and abandoned it to be members of Rust

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

forums. That said, our survey population comprised people with a variety of general and Rust-specific development experience. 26% of respondents had used Rust for less than one year and they often held similar opinions to more experienced Rust users. Our results uncovered a wide variety of specific challenges and benefits which can provide novel insights into the human factors of secure language adoption.

Participants largely perceived Rust to succeed at its goals of security and performance. Other key strengths identified by participants include an active community, high-quality documentation, and clear error messages, all of which make it easy to find solutions to problems. Further, participants indicated that overall Rust benefits the development cycle in both speed and quality, and using Rust improved their mental models of secure programming in ways that extend to other languages.

However, participants also noted key drawbacks that can inhibit adoption, most seriously a steep learning curve to adjust to the paradigms that enforce security guarantees. Other concerns included dependency bloat, limited library support, slow compile times, high up-front costs, worries about future stability and maintenance, and apprehension about the ability to hire Rust programmers going forward. For our participants, these negatives, while important, were generally outweighed by the positive aspects of the language.

Lastly, participants offered advice for others wanting to advocate adoption of Rust or other secure languages: be patient, pick projects playing to the language's strengths, and offer support and mentorship during the transition.

Analyzing our findings, we offer recommendations aimed at supporting greater adoption of Rust in particular and secure languages generally. The popularity of Rust with our participants highlights the importance of the ecosystem — tooling, documentation, community — when developing secure languages and tools that users will actually want to use. Our results also suggest that perhaps the most critical path toward increased adoption of Rust in particular is to flatten its learning curve, perhaps by finding ways to gradually train developers to use Rust's ownership and lifetimes. Further, we find that much of the cost of adoption occurs up front, while benefits tend to accrue later and with more uncertainty; security advocates should look for ways to rebalance this calculus by investing in a pipeline of trained developers and contributing to the longevity and stability of the Rust ecosystem.

2 Background

Rust is an open-source systems programming language created by Mozilla, with its first stable release in 2014. Rust's creators promote its ability to "help developers create fast, secure applications" and argue that Rust "prevents segmentation faults and guarantees thread safety." This section presents Rust's basic setup and how it aims to achieve these benefits. For those with a deeper interest, we recommend the tutorial offered in the official Rust Programming Language Book [21].

2.1 Core features and ecosystem

Rust is a multi-paradigm language, with elements drawn from functional, imperative, and object oriented languages. Rust's **traits** abstract behavior that types can have in common, similarly to interfaces in Java of typeclasses in Haskell. Traits can be applied to any type, and types need not specifically mention them in their definitions. Objects can be encoded using traits and structures. Rust also supports **generics** and **modules**, and a sophisticated **macro system**. Rust's variables are **immutable by default**: once a value is bound to a variable, the variable cannot be changed unless it is specifically annotated as mutable. Immutability eases safe code composition, and plays well with ownership, described shortly. Rust also enjoys **local type inference**: types on local variables are generally optional, and can be inferred from their initializer. Rust also supports **tagged unions** ("enums") and **pattern matching**, which allow it to, for example, avoid the need for a *null* value (the "billion dollar mistake" [19]).

Rust has an integrated build system and package manager called **Cargo**, which downloads library packages, called **crates**, as needed, during builds. Rust has an official community package registry called crates.io. At the time of writing, Crates.io lists more than 49,000 crates.

2.2 Ownership and Lifetimes

To avoid dangerous, security-relevant errors involving references, Rust enforces a programming discipline involving ownership, borrowing, and lifetimes.

Ownership. Most type-safe languages use garbage collection to prevent the possibility of using a pointer after its memory has been freed. Rust prevents this without garbage collection by enforcing a strict *ownership*-based programming discipline involving three rules, enforced by the compiler:

1. Each value in Rust has a variable that is its ***owner***.
2. There can only be ***one owner at a time*** for each value.
3. A value is ***dropped*** when its ***owner goes out of scope***.

An example of these rules can be seen in Listing 1. In this example, `a` is the owner of the value "example." The scope of `a` starts when `a` is created on line 3. The scope of `a` ends on line 5, so the value of `a` is then dropped. In the second block of code, `x` is the initial owner of the value "example." Ownership is then transferred to `y` on line 11, which is why the print on line 13 fails. The value cannot have two owners.

Borrowing. Since Rust does not allow values to have more than one owner, a non-owner wanting to use the value must *borrow* a reference to a value. A borrow may take place so long as the following invariant is maintained: There can be (a) just one mutable reference to a value `x`, *or* (b) any number of

```
1  {
2  //make a mutable string and store it in a
3  let mut a = String::from("example");
4  a.push_str("␣,␣text"); //append to a
5  }
6  //scope is now over so a's data is dropped
7
8  {
9  //make a mutable string and store it in x
10 let x = String::from("example");
11 let y = x; //moved ownership to y
12 println!("y␣is␣{}", y); //allowed
13 println!("x␣is␣{}", x); //fails
14 }
```

Listing 1: Examples of how ownership works in Rust

immutable references to `x` (but not both). An example of the rules of borrowing can be seen in Listing 2. In this example, a mutable string is stored in `x`. Then, immutable references are made ("borrowed") on lines 6 and 8. However, the attempt to mutate the value on line 10 fails since `x` cannot be mutated while it has borrowed (immutable) references. Line 12 fails in the attempt to make a mutable reference: `x` cannot have both a mutable and an immutable reference. Once we reach line 13, the immutable references to `x` have gone out of scope and been dropped. `x` is once again the owner of the value and possesses a mutable reference to the value, so line 14 does not fail. In the second code block, starting on line 15, a mutable reference is made to the value. Attempts to make a second mutable reference on lines 19 and 21 fail because only one mutable reference can be made to a value at a time.

The ownership and borrowing rules are enforced by a part of the Rust compiler called the *borrow checker*. By enforcing these rules the borrow checker prevents vulnerabilities common to memory management in C/C++. In particular, these rules prevent dangling pointer dereferences and double-frees (only a sole, mutable reference may be freed), and data races (a data race requires two references, one mutable).

Unfortunately, these rules also prevent programmers from creating their own doubly-linked lists and graph data structures. To create complex data structures, Rust programmers must rely on libraries that employ aliasing internally. These libraries do so by breaking the rules of ownership, using unsafe blocks (explained below). The assumption is that libraries are well-vetted, and Rust programmers can treat them as safe.

Lifetimes. In a language like C or C++, it is possible to have the following scenario: (1) you acquire a resource; (2) you lend a reference to the resource; (3) you are done with the resource, so you deallocate it; (4) the lent reference to the resource is used. Rust prevents this scenario using a concept called *lifetimes*. A lifetime names a scope, and a lifetime annotation on a reference tells the compiler the reference is valid only within that scope. For example, the lifetime of variable `a` in Listing 1 ends on line 5 where the scope of `a`

```
1  {
2  //make a mutable string and store it in x
3  let mut x = String::from("example");
4     {
5     //make immutable reference to x
6     let y = &x; //allowed
7     //make second immutable reference to x
8     let z = &x; //alllowed
9     println!("x␣is␣{}.␣y␣is␣{}", x, y) //allowed
10    x.push_str("␣,␣text"); //fails
11    //make mutable reference to x
12    let mut a = &mut x; //fails
13    } //drops y and z; x owner again
14 x.push_str("␣,␣text"); //allowed
15    {
16    //make mutable reference to x
17    let mut a = &mut x; //allowed
18    a.push_str("␣,␣text"); //allowed
19    x.push_str("␣,␣text"); //fails
20    //make second mutable reference to x
21    let mut b = &mut x; //fails
22    } //drops a; x is owner again
23 }
```

Listing 2: Examples of how borrowing works in Rust

ends. Similarly, the lifetime of `a` in Listing 2 ends on line 22.

2.3 Unsafe Rust

Since the memory guarantees of Rust can cause it to be conservative and restrictive, Rust provides escape hatches that permit developers to deactivate some, but not all, of the borrow checker and other Rust safety checks. We use the term *unsafe blocks* to refer generally to unsafe Rust features. Unsafe blocks allows the developer to:

- Dereference a raw pointer
- Call an unsafe function or method
- Access or modify a mutable global variable
- Implement an unsafe trait
- Access a field of a union

Unsafe functions and methods are not safe in all cases or for all possible inputs. Unsafe functions and methods can also refer to code that a developer wants to call that is in another language. Unsafe traits refer to traits with at least one unsafe variant. Lastly, unions are like structs but only one field in a union is used at a time. To use unsafe blocks, the relevant code construct is labeled with keyword `unsafe`.

3 Method

To understand the benefits and drawbacks to adopting Rust, we conducted semi-structured interviews with senior and professional software engineers working at technology companies who were using Rust or attempting to get Rust adopted. To examine the resulting findings in a broader ecosystem, we then distributed a survey to the Rust community through

Sect.	Description and Example Questions
1	Technical Background (General, and Rust) • How long have you been programming? • How long have you been programming in Rust?
2	Learning and using Rust • How easy or difficult did you find Rust to learn? • How would you rate the quality of available Rust docs? • When I encounter a problem or error while working in Rust, I can easily find a solution to my problem?
3	Work (general), and using Rust for work • Did anyone at your employer have apprehensions about using Rust? • What one piece of advice would you give to someone who is trying to get Rust adopted?
4	Comparing Rust to other familiar languages • How would you rate the quality of Rust compiler and run-time error messages compared to [*chosen language*]?
5	Rust likes/dislikes & unsafe blocks • Which of the following describes your use of unsafe blocks while programming in Rust?
6	Porting and interoperating with legacy code • What language(s) have you ported from?
7	Demographics about participants • Please select your highest completed education level

Table 1: Survey sections and example questions.

various online platforms.

3.1 Interviews and Surveys

Interview protocol. From February through June 2020, we conducted 16 semi-structured interviews via video-conferencing software.

Each interview included two phases. Phase one asked the participants how they discovered and learned Rust, as well as about instances when they or their company decided to use Rust for projects (or not) and why. In the second phase, we asked more technical questions about programming with Rust, including what features of the language/ecosystem they like/dislike compared to other familiar languages, as well as opinions about features relating to Rust's security, such as ownership and unsafe blocks.

Each session lasted about an hour, giving participants a chance to share detailed experiences. The full interview protocol is given in Appendix A.

Survey. The survey was designed to mirror the interviews, with closed-item answer choices inspired by answers from the open-ended interview questions. The survey was broken into seven sections; Table 1 tabulates the sections and provides some example questions. The full survey is given in Appendix B. It was active from July to September 2020.

Recruitment. To recruit for the interviews, we contacted a longtime member of the core Rust team and asked them to connect us with software engineers who were active members or leaders of teams using or adopting Rust at their employers. From these initial referrals, we snowball-sampled more interviewees (asked participants to refer us to peers). We also recruited participants referred to us by colleagues, and contacted people and companies quoted or listed on the Rust website [35]. We focused on recruiting participants with senior, leadership, or other heavily involved roles in the Rust adoption process. We interviewed participants until we stopped hearing substantially new ideas, resulting in a total of 16 participants. This sample size aligns with qualitative best practices [16].

To recruit participants for the survey, we advertised on several Rust forums and chat channels: Reddit channel `r/rust`; Rust Discord community channels `embedded`, `games-and-graphics`, `os-dev`, `gui-and-ui`, and `science-and-ai`; Rust Slack beginners channel; Rust Facebook Group; and the official Rust Users Forum. Those who wanted to participate were directed to follow a link in the notice that took them directly to the survey.

Ethics. Both the interview and the survey were approved by University of Maryland's ethics review board. We obtained informed consent before the interview and the survey. Given that we were asking questions about their specific companies and the work they were doing, participants were informed that we would not disclose the specific company they worked for. They were reminded that they could skip a question or stop the interview or survey at any time if they felt uncomfortable.

3.2 Data analysis

Once the interviews were complete, two team members transcribed the audio recordings and then analyzed them using iterative open coding [8]. The interviewer and the other team member independently coded the interviews one at a time, developing the codebook incrementally and resolving disagreements after every transcript. This process continued until a reasonable level of inter-rater reliability was reached measured with the Krippendorff's α statistic [22]. After seven interviews, the two researchers achieved a Krippendorff's $\alpha = 0.80$, calculated using ReCal2 [11]. This level of agreement is above the commonly recommended thresholds of 0.667 [18] or 0.70 [23] for exploratory research and meets the more general minimum threshold recommended by Krippendorff of 0.8 [22]. Once a reliable codebook was established, the remaining nine interviews were evenly divided among the two researchers and coded separately.

We report the results of our closed-response survey questions using descriptive statistics. While we did not have any questions as specific attention checks, we evaluated the responses for completeness to ensure that we removed all low-quality responses. We did not remove many responses as can be seen in Section 4. Since our work is exploratory, we did

not have any hypotheses, so we do not make any statistical comparisons. Free response questions from the survey were analyzed by one researcher using the same codebook from the interview. When new codes were added to the codebook, they were back-applied to the interviews.

Throughout the following sections, we use *I* to indicate how many interview participants' answers match a given statement, and use *S* to denote how many survey participants' answers do, either as a percentage (closed-item questions) or count (open-ended ones). We report on interview and survey results together, as the results generally align. We report participant counts from the interviews and open-ended items for context, but not to indicate broader prevalence. If a participant did not voice a particular opinion, it does not necessarily mean they disagreed with it; they simply may not have mentioned it.

3.3 Limitations

Our goal with the interviews was to recruit people who had substantial experience, and preferably a leadership role, in attempting to adopt Rust at a company or team. We believe we reached the intended population. Only one interviewee failed to see Rust adopted at their employer, but all interviewees faced similar adoption challenges.

For the surveys, our goal was to reach a broad variety of developers with a range of Rust experiences, in order to capture the widest range of benefits and drawbacks. We did reach participants with a wide range of Rust experience, in part because we targeted many Rust forums, including some specifically for beginners. However, because all of these forums are about Rust, we may not have reached people who have tried Rust but abandoned it, or those who considered it but decided against it after considering potential pros and cons. In addition, these forums are likely to overrepresent Rust enthusiasts compared to those who use the language because they are required to. Further, there could be self-selection bias: because we stated our goal of exploring barriers and benefits to adopting Rust when recruiting, those with particular interest in getting Rust adopted may have been more likely to respond.

Taken together, these limitations on our survey population suggest that our results may to some extent overstate Rust's benefits or may miss some drawbacks that drive people away from the language entirely. Nonetheless, our results uncovered a wide variety of challenges and benefits that provide novel insights into the human factors of secure language adoption. Given the general difficulty of recruiting software developers [36], and the particular difficulty of reaching this specific subpopulation, we consider our sample sufficient.

4 Participants

Interview participants. We interviewed 16 people who were active members of teams using or adopting Rust at their company. Our participants mostly held titles related to software development (I = 12) and worked at large technology companies (more than 1000 employees, I = 9), as shown in Table 1 in Appendix C. Most of them had worked in software development for many years and were members of, several leading, teams building substantial project(s) in Rust at their employers. Many were Rust evangelists at their companies. Their companies develop social media platforms, bioinformatics software, embedded systems, cloud software, operating systems, desktop software, networking software, software for or as research, and cryptocurrencies.

Survey respondents. We received 203 responses to our survey. We discarded 25 (12%) incomplete surveys, which left 178 complete responses. Respondents were predominantly male (88%), young (57% below the age of 30 and 88% below the age of 40), and educated (40% had a bachelor's degree and 28% had a graduate degree). Our participants were relatively experienced programmers (53% had more than 10 years of programming experience and 85% had at least 5 years). Seventy-two percent of our participants were currently employed in the software engineering field and worked in a variety of application areas, as shown in Table 2 in Appendix C. Additionally, our participants had used a variety of languages in the prior year, as shown in Figure 1.

Our survey participants were fairly experienced at programming in Rust (37% had been programming in Rust at least 2 years and 74% at least 1 year). Ninety-three percent of respondents had written at least 1000 lines of code and 49% had written at least 10,000 lines of code. Forty-six percent had only used Rust for a hobby or project, 2% had only used it in a class, 14% had maintained a body of Rust code, and 38% had been paid to write Rust code. Most of our respondents were currently using Rust (93%), while some had used it on projects in the past but were not currently using it (7%). While our survey participants had a broad variety of experiences, they may underrepresent people who tried and turned away from Rust—such people would probably not be members of the Rust forums in which we advertised. As such, our results may overstate Rust's benefits or may miss some drawbacks that drive people away from the language entirely. Nevertheless, we believe our results offer practical insights relevant to the adoption of secure programming languages.

Survey respondents' companies. Nearly half of our respondents were using Rust for work (49%). Of those using Rust for work, most were using Rust as a part of a company or large organization (84%), rather than as a freelance assignment. We gathered further details about these 87 respondents' companies. They were primarily small (53% of the 87 worked for companies with 100 or fewer employees and 74% worked for a company with less than 1000 employees). They mostly developed alone (50%) or in small teams of two to five people (40%) at their companies, and their companies had legacy codebases of varying sizes (88% had 500,000,000 or fewer

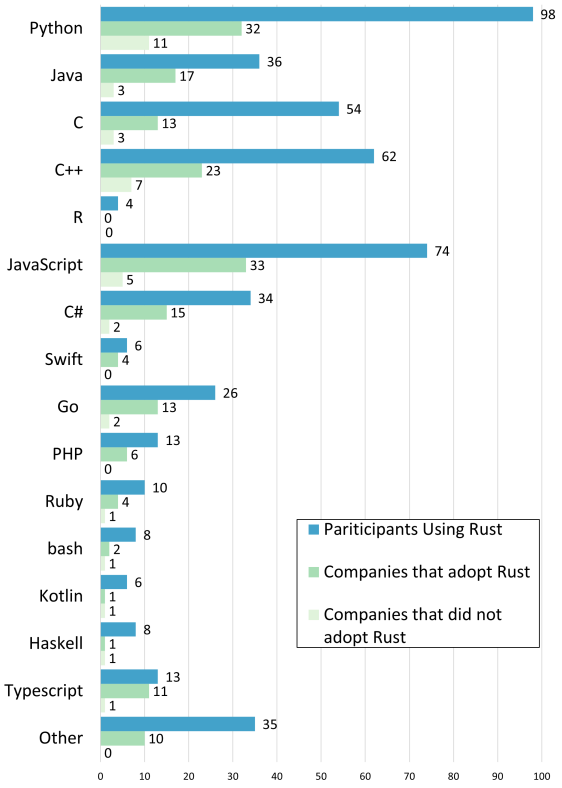


Figure 1: Languages used in the past year by survey participants (counts), companies that had adopted Rust, and companies that considered but didn't adopt Rust. Ordered according to the IEEE 2019 top programming languages list [20].

lines of code and 64% had 1,000,000 or fewer lines of code). A variety of languages were used at respondents' companies (whether they had adopted Rust or not), as shown in Figure 1.

5 How is Rust being used?

This section and the next two analyze our interview and survey results. We first examine how our participants are using Rust.

5.1 Applications

Interview participants reported using Rust in a variety of application areas, including databases (I = 3); low-level systems such as operating systems, device drivers, virtual machine management systems, and kernel applications (I = 5); data processing pipelines (I = 1); software development applications such as monitoring resource usage (I=2); and compilers and programming languages tools (I = 2).

Participants did not always consider Rust the best tool for the job. When asked to select application areas for which Rust is not a strong fit, they most frequently mentioned mobile (I = 1, S = 44%), GUI (I = 3, S = 37%), and web applications (I = 3, S = 17%). For example, I9 said "*Strongly typed languages like*

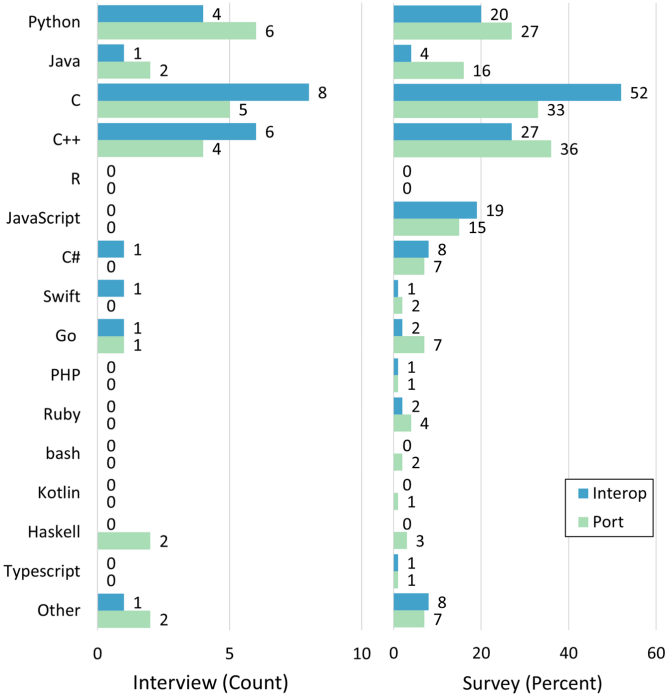


Figure 2: Interview and survey participants porting (survey n = 123) to Rust from and interoperating (survey n = 84) Rust with other languages. Languages are ordered via ranking on the IEEE 2019 top programming languages list [20].

Rust... lend themselves much more to systems programs... and less to web applications and things that you want to be very flexible." Interestingly, 13 survey participants who selected web development as a bad fit for Rust also chose web development as one of the things they do for work. Several participants mentioned that Rust is a poor fit for prototyping or one-off code (I = 6, S = 3%). I4 explained, "*I still prototype everything in C++ because it just works faster ... [Rust's] not a great prototyping language.*"

5.2 Porting and interoperating

Because Rust is relatively new, using it often requires porting or interoperating with legacy code written in other languages.

Most participants had ported code from another language into Rust (I = 14, S = 69%). They had ported from a variety of languages (Figure 2). Interview participants found porting code from Python to be easy (I = 5); similarly, where 70% of survey respondents who had ported from Python (n=33) found it either somewhat or extremely easy. In contrast, fewer participants found porting from C (I = 2; S = 54%, n = 41) and C++ (I = 2; S = 52%, n = 44) somewhat or extremely easy. I11 said porting from C++ is "*much harder because... you structure your data with movability [mutability].*"

Many participants had written code to interoperate with Rust (I = 13, S = 47%), starting from a variety of languages

(Figure 2). Ease of interoperation varied by language somewhat differently than ease of porting. Almost three-quarters of participants who had interoperated with C found it at least somewhat easy (I = 6; S = 70%, n = 44). A majority also rated Python somewhat or extremely easy (I = 2; S = 53%, n = 17). Less than half considered C++ at least somewhat easy (I = 2; S = 43%, n = 23). I6 attributes this to the fact that "*the C++ side is just the Wild West. There's rampant aliasing . . . and none of that is going to play by Rust's rules.*"

5.3 Unsafe blocks

As described in Section 2, unsafe blocks allow the programmer to sidestep borrow-checking, which can be too restrictive in some cases. Because unsafe blocks may potentially compromise Rust's safety guarantees, we investigate how they are used and what if any error-mitigation strategies exist.

Unsafe blocks are common and have a variety of uses. Most participants had used unsafe blocks (I = 15, S = 72%). Use-cases included foreign-function interfacing (I = 11, S = 70%), increasing code performance (I = 3, S = 40%), kernel-level interaction (I = 1, S = 35%), hardware interaction (I = 4, S = 34%), and memory management (I = 4, S = 28%). For example, I14 uses unsafe blocks to "*wrap all of our. . . code for accessing hardware,*" since they had to do things like "*write values into this offset relative to the base address register,*" which is prohibited by Rust ownership rules.

Few companies have unsafe-code reviews. To avoid introducing problems Rust otherwise guarantees against, companies may implement a procedure to check that "unsafe" code is actually safe. However, unsafe-review policies were uncommon at our participants' employers (I = 7, S = 28%). Where specific policies do exist, the most common approach is a thorough code review (I = 2, S = 68%). For example, at S118's company, the review policy is "*pretty simple: pay extra close attention to unsafe blocks during code review.*" To help code reviewers, developers use comments to explain why the code is actually safe (I = 5, S = 21%). I5 commented, "*I guess the only formal thing is that every unsafe block should have a comment saying why it is in fact safe.*" These comments may aim to explain important safety invariants [3].

6 Benefits and drawbacks of Rust

This section explores benefits and drawbacks of Rust related to technical aspects of the language, learning the language, the Rust ecosystem, and Rust's effect on development.

6.1 Technical benefits of Rust

Participants largely are motivated by, and agree with, Rust's claims of performance and safety [34].

Safety is important. Many participants identified Rust's safety assurances as benefits. They listed memory safety (I = 10, S = 90%), concurrency safety (I = 6, S = 84%), immutability by default (I = 4, S = 74%), no null pointers (I = 3, S = 81%), Rust's ownership model (I = 2, S = 75%), and lifetimes (I = 2, S = 55%). As I5 said about Rust's strengths, "*The safety guarantees, like 100%.... That's why I use it. That's why I was able to convince my boss to use it.*"

So is performance. Participants were also drawn to Rust's promise of high performance. Respondents explicitly listed performance (I = 7, S = 87%) and, less explicitly, lack of garbage collection (I = 3, S = 63%) as reasons to like the language. I1 describes the appeal of Rust: "*it gives you the trifecta of performance, productivity, and safety.*"

6.2 Learning Rust: Curiosity vs. reality

We next review participants' experiences learning Rust.

Most chose to learn Rust because it is interesting or marketable. Most participants selected, as their primary reason(s) to learn Rust, curiosity (I = 2, S = 90%). Other said they had heard about it online or it was suggested by a friend (I = 12, S = 25%). Participants also believed knowing Rust was a marketable or useful job skill (I = 7, S = 22%).

Rust is hard to learn. Possibly the biggest drawback of Rust is its learning curve. Most participants found Rust more difficult to learn than other languages (I = 7, S = 59%). I14 said Rust has "*a near-vertical learning curve.*"

Asked how long it took to learn to write a compilable program without frequently resorting to the use of unsafe blocks, a plurality of participants said one week to one month (I = 2, S = 41%), less than one week (I = 0, S = 27%), or one to six months (I = 3, S = 25%). Notably, six survey participants were not yet able to do this. Interviewees had similar experiences. I3 "*didn't feel fully comfortable with Rust until about three months in, and really solid programming without constantly looking stuff up until about like six months in.*" Five interviewees said it takes longer to get Rust code to compile than another language they are comfortable with. Survey participants agreed (S = 55%, Figure 3). S161 commented, "*You spend 3–6 months in a cave, breathing, eating and sleeping Rust. Then find like-minded advocates who are prepared to sacrifice their first born to perpetuate the unfortunate sentiment that Rust is the future, while spending hours/days/weeks getting a program to compile and run what would take many other 'lesser' languages a fraction of the time.*"

The borrow checker and programming paradigms are the hardest to learn. Seven interviewees reported that the biggest challenges in learning Rust were the borrow checker and the overall shift in programming paradigm. A few survey participants (S = 3) noted this in free-response as something they explicitly did not like about Rust. S136 did not like "*having to redesign code that you know is safe, but the compiler*

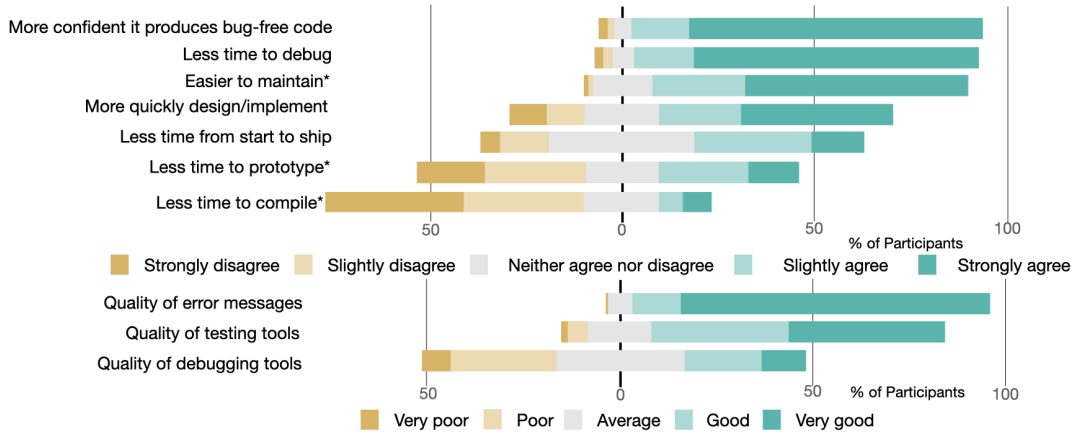


Figure 3: Likert-style responses comparing Rust to a language survey participants were most comfortable with. Green bars advantage Rust; gold bars advantage the other language. Questions with a * have been flipped in polarity for consistency.

doesn't." I8 echoes this frustration: "*There are new paradigms that Rust sort of needs to teach the programmer before they can become super proficient. And that just makes the learning curve a little bit higher, and that did frustrate a number of people, . . . because it's something that's sufficiently different from other things they're used to.*" This could pose a problem for adoption, if the frustration of learning these new paradigms turns developers away from Rust altogether.

6.3 Rust ecosystem: Good and getting better

The Rust ecosystem influences organizational adoption, because it provides needed support for large projects. Participants identified a variety of current benefits and drawbacks.

Tools are easy to use and well supported, but slow. Asked how Rust's tooling compared to the other language they were most comfortable programming in, 75% of survey participants found it either very good or good (Figure 3). Also in Figure 3, most survey participants found Rust's compiler and runtime error messages to be good or very good compared to their reference language (S = 92%). Beginners (less than one year of experience, n = 47) felt this way, too (S = 87%). A large majority (I = 8, S = 97%) listed the compiler's descriptive error messages as a major problem-solving benefit. I9 comments: "*Most of the time the compiler is very, very good at telling you exactly what the problem is.*" When it doesn't, "*Rust is an exercise in pair programming with the compiler,*" wrote S176. Participants also liked the crates ecosystem (I = 4, S = 83%). For example, I7 said, "*I also just love the cargo tooling; it's so easy to get crates.*" While participants like the tooling, they dislike that Rust has a long build time (I = 4, S = 55%). I16 said, "*Compile times are pretty bad. . . I don't think Rust will ever get close to like Go level of compile speed.*"

Easy to find solutions. Despite the challenging learning curve, participants report it is easy to find solutions to problems they encounter when developing in Rust (I = 14, S = 79% overall, 70% of beginners). Participants attribute this to good compiler errors (discussed above), good official documentation (I = 3, S = 91%), and the helpfulness of the Rust developer community, in-person and online (I = 5, S = 46%). I5 notes the "*very accessible documentation and kind of an active community. . . on Stack Overflow and so forth. I feel like if I have a problem with Rust, I Google it and there's always an answer.*"

Rust lacks libraries and infrastructure and causes dependency bloat. Despite the high quality of available tools and libraries, Rust still lacks some critical libraries and infrastructure, perhaps in part because it is fairly new. When asked what they dislike about the language, many participants noted the lack of available libraries (I = 3, S = 39%). I4 agrees:"*It feels like you're reinventing a lot of infrastructure, right? So, I've felt that it's slower [to develop with].*" Additionally, participants complained about a tendency toward dependency bloat (I = 4, S = 34%). I4 agrees: "*You know (cargo) goes and pulls every dependency ever. . . That part's bad. It encourages dependency bloat, which, in a security focused area, is also the exact opposite of what you want.*"

6.4 Mostly positive impact on development

We find Rust offers development-cycle benefits that may in part offset its learning curve and upfront adoption costs.

Rust improves confidence in code. A key benefit mentioned by participants is that once Rust code compiles, developers can be fairly confident that the code is safe and correct. Four interview participants mentioned that they spend less time debugging in Rust than in other languages; this was supported in the survey, when 89% of respondents (87% of beginners) slightly or strongly agreed they spend less time debugging compiled Rust code than code in another language they are comfortable with. Interviewees also mentioned that

Rust makes them more confident their production code is bug-free (I = 9); 90% of survey respondents slightly or strongly agreed. I16 said, "*The thing that I like the most about Rust overall is the fact that if the compiler is okay with your code then it will probably mostly be working.*"

Rust improves productivity in the development cycle. While the initial time to design and develop a solution in Rust is sometimes long and/or hard to estimate due to unforeseen conflicts with the borrow checker, interview participants felt — and survey participants agreed or strongly agreed — that Rust reduced development time overall, from the start of a project to shipping it, compared to other languages they were comfortable with (I = 7, S = 45%). I1 said, "*They see how well these projects go in comparison to the C++ projects;... and they've seen quantitatively that the Rust projects they've been working on have been a dramatically better experience and more predictable and a faster lifecycle.*"

Additionally, five interview participants noted that they could more quickly design and implement bug-free code in Rust than in another language they were comfortable with. This is echoed in the survey, where 61% of participants agreed or strongly agreed, as shown in Figure 3. 81% of survey participants also strongly or slightly disagreed that maintaining code is more difficult in Rust than in other languages. The reported improvement in developer productivity and code quality resulting from the use of Rust means that companies and organizations can ship better-quality code in less time.

Rust improves safe development in other languages. Most participants report Rust has had at least a minor positive effect on their development in another language they're comfortable with (I = 10, S = 88%). These participants said Rust causes them to think about ownership (I = 5, S = 68% of 155), data structure organization (I = 6, S = 59%), use of immutability (I = 0, S = 48%), iteration patterns (I = 1, S = 45%), memory lifetimes (I = 4, S = 37%), and aliasing (I = 2, S = 25%). This is encouraging, as it shows developers carry over the safety paradigms that they are forced to consider in Rust when working in other languages. This is exemplified by S40, who said, "*Once you learn Rust, you are one with the borrow checker — it never leaves you. I now see many of the unsafe things I have been doing in other languages for years, (but probably not all of them, as I am human and not a compiler).*"

Notably, a few participants volunteered that Rust has even made them stop using C++ altogether (I = 2, S = 2). S26 said, "*It has made me stop working with C++. I really do feel that Rust replaces C++'s use cases well.*"

Overall, these results hint that the high cost of learning Rust can be worth it, providing longer-term benefits in other applications. This means developers may write more secure code in other languages, and organizations may get benefit out of investing time in their developers learning Rust.

7 Organizational adoption of Rust

While the Rust benefits identified by our participants may also apply to organizations, many participants mentioned experiencing pushback from teammates or managers (I = 9, S = 41%). Notably, some participants' attempts at adoption were unsuccessful (I = 1, S = 20%).

Participants identified several organization-level apprehensions about adopting Rust. We divide these into two categories: those that may apply to any change in programming language, and those that are specific to Rust. Rust-specific concerns closely mirror the drawbacks of adopting Rust individually our participants identified above.

7.1 Apprehensions about any new language

Unfamiliarity with the language. Many participants cited unfamiliarity with Rust as one reason people were worried about adopting or did not adopt Rust at their company (I = 2, S = 69%). Any change to an unfamiliar language could create uncertainty or apprehension.

Avoiding unnecessary changes. Participants also reported a general desire to avoid unnecessary change. In particular, several participants' companies are reluctant to add any new languages (I = 2, S = 46%). As I2 explained, "*Not wanting to have too many languages in play at the company simultaneously, and so just a general conservatism there around not wanting to pick up new languages willy nilly.*"

Business pressures. Some participants said their companies were concerned about using or did not want to use Rust because there was time pressure to deliver a product, and they did not want to invest the time to get a new language and its infrastructure up and running (I = 1, S = 38%).

Lack of fit with existing codebase and ecosystem. Participants reported that lack of compatibility with the existing development ecosystem (I = 2, S = 27%) or interoperability with the existing codebase (I = 4, S = 27%) were concerns for their employers. As I6 said, "*It's very different for your developers or your managers who are managing a large, mature C++ ecosystem. They're much more skeptical of Rust. Maybe not on its merits, but just in the practical terms of how do I integrate this with my huge existing ecosystem?*"

7.2 Apprehensions specific to Rust

Rust's steep learning curve. The difficulty of learning Rust was among the biggest concerns participants encountered at their companies (I = 3, S = 50%). I14 said one worry was "*how are we going to ramp programmers up? And I think Rust in particular has this reputation of having a very steep learning curve.*" Some participants' companies were also concerned about potential reduction in developer productivity (I

= 3, S = 29%) or difficulty maintaining Rust code (I = 1, S = 23%). At I7's company, for example, "*the main concern was that it would be taking too long to use Rust.*"

Rust's maturity and maintenance. Since Rust is relatively new, some participants cited company concerns about the maturity and maintenance of its tooling and ecosystem, as well as whether it would be around long-term (I = 4, S = 29%). As I1 said, "*If [developers are] launching a new codebase in a new language, it's going to take them a year, maybe three years, to develop the things, and they care where the ecosystem will be at that point.*" Other comments reflected a lack of trust in the Rust toolbase (I = 3, S = 8%). I14's company worried, "*How well supported is the tool chain? How mature is the compiler? . . . Rust is a new language.*"

Difficulty hiring Rust developers. Stemming possibly from the newness and the difficulty of learning Rust, some participants reported their companies worried about the ability to hire Rust developers (I = 5, S = 42%). I11, for example, said, "*Do we really want to keep this thing in Rust? It's hard to find a new person for the team. . . because we don't have . . . a huge pool of Rust programmers.*"

7.3 Ways to encourage adoption

Despite these apparent apprehensions, many participants' companies still adopted Rust (I = 15, S = 49%). We report their suggestions for enabling adoption.

Pick projects carefully. Participants suggest that advocates pick initial projects for Rust carefully. Projects should fit Rust's strengths (I = 5, S = 2), both in terms of language design and available tooling. S62 recommended, "*Pick projects that are suitable for Rust, based on how mature the ecosystem (crates) is at supporting that type of project.*" S99 similarly commented, "*Don't try to port paradigms or design patterns from other languages.*" Participants also advise starting small (I = 5, S = 12). S95 said, "*Start small. There are many little problems that Rust programs solve well, which builds trust.*"

Demonstrate value. Participants argue that adoption hinges on demonstrating the value of using Rust. Most importantly, participants said advocates must argue that Rust offers a measurable improvement over the company's current language (I = 6, S = 10). While Rust touts its guarantees for safety and correctness, companies want to know the time and effort they allot to tackle the Rust learning curve will result in a major benefit. For example, S65 suggests, "*If you give a presentation about Rust, focus on concepts unique to Rust and what they offer; what matters is the idea that somehow it's possible to write safe, concurrent & fast software thanks to those concepts.*" This echoes results from Haney et al. suggesting security advocates must demonstrate value to motivate people to take appropriate security actions [17].

Participants emphasize being clear and straightforward about Rust's drawbacks, while arguing that the benefits outweigh them. I3 recommends "*rewrit[ing] [code] in Rust and swap[ping] it in.. . . And then you say look, this provides the same API. You didn't even know.*" If advocates can show their managers and teammates that using Rust had no negative effect on the codebase, they may be less apprehensive about its effect on productivity and timelines. Other participants recommend using a prototype to show that Rust is worth adopting (I = 4, S = 4). As I11 said, "*We're gonna do a prototype. If doesn't work we'll just kill it*"

Account for upfront costs. Another strategy suggested by participants is to be clear about, and attempt to mediate, upfront costs, including additional time to design for the ownership paradigm as well as challenges related to tooling and dependencies (all discussed above). Participants suggest advocates spend time significant time planning tooling (I = 4, S = 2). I14 specifically advised to "*invest in your tooling upfront. Everybody starts out with Cargo, and Cargo is wonderful for what it does, but it has problems.*" Due to the steep learning curve, participants also suggest that advocates budget enough time to get started (I = 3, S = 1). For example, I5 advised, "*Factor in the learning curve and ramping up period that you're going to need to do. Because with initial adoption, you're probably not going to be able to hire like Rust programmers . . . for a decent size project, and . . . it does take a long time to kind of become productive in Rust, especially compared to some other languages, but if you are expecting that then over the long term you're gonna get big advantages.*"

Be helpful and have a good support system. Given the steep learning curve, participants emphasize the need for advocates to be willing and able to help new developers (I = 4, S = 2). They recommend the advocate themselves be a knowledgeable Rust developer (I = 2, S = 8): S118 suggests "*Make yourself an expert (e.g., via personal projects and study) and share your expertise generously. People will feel more comfortable with an unfamiliar language if they have a friendly, helpful expert on their team. Finally, be patient. . . . Being friendly, helpful, and humble usually works better than being pushy, righteous, and evangelical.*" Similarly, S73 said, "*Make sure you are willing to mentor aggressively for a long time.*" Further, some participants suggest a formal support system for teaching and mentoring new Rust developers (I = 3, S = 3). I8 advised, "*Try to just have some good support for newer engineers.. . . If you do happen to have a couple engineers who are more proficient in Rust and are willing to help, . . . have those engineers help the newer ones.*"

Be persistent but patient. Companies may not always buy in to adopting Rust immediately. Some participants suggest advocates for Rust be persistent (I = 1, S = 3). S84 suggested adopters should "*keep at it and try to get coworkers to pick it up as well. Strength in numbers.*" However, participants also suggested advocates be patient (I = 3, S = 5) and not

"*expect [their employer] to agree to making any changes at first.*" Advocates need to "*give Rust time and be patient, the memory model and lack of OOP combine to make it difficult for existing programmers to jump into.*" This advice — which ties into the steep learning curve and lack of language maturity discussed in Section 7.2 above — aligns well with Haney et al.'s finding that building relationships and trust helps with the adoption of secure systems and technologies [17].

8 Discussion and recommendations

Our results demonstrate that there are drawbacks to adopting Rust but, at least for our participants (many of whom are Rust enthusiasts), the benefits appear to outweigh them. This section summarizes what we can learn from Rust's success to date, and recommends steps toward improving adoption or use of Rust itself, as well as other secure languages and tools.

Making secure tools and languages appealing. All but one survey respondent said they would either probably or definitely use Rust again in the future (S = 99%) and many survey participants felt that their employer would likely use Rust again (S = 88%). This mirrors the results of the Stack Overflow Developer Survey, where Rust has been the "most loved" language for the last five years in a row [46].

Our results shed some light on why this might be. We confirmed that to a large extent, Rust is perceived to meet its motivating goals of security and performance. Further, Rust's tools provide high-quality feedback (e.g., error messages), the language boasts good documentation, and it has an active and helpful online community; all of these were deemed important in prior studies of language adoption [25]. Good documentation and a responsible and attentive community are also known to be important for encouraging adoption of secure APIs and programming patterns [1, 2].

Flatten the learning curve. Participants overwhelmingly report that learning Rust had a positive effect on their development skills in other languages, including by internalizing memory safety-relevant concepts such as ownership and lifetimes. Rust caused participants to shift their programming mental models, which echoes prior work showing that "mindshifts are required when switching paradigms" [45].

Unfortunately, our participants also report that Rust can be very difficult to learn (Section 6.2) precisely because of the difficulty of adhering to these concepts (as enforced by Rust's ownership and lifetime rules). As observed with other security tools, Rust's learning curve may be turning some developers and/or organizations away from using it [48].

Finding ways to flatten this curve could have a big impact. For example, it may make sense to develop a version of Rust that allows users to incrementally learn the difficult concepts of ownership and borrow checking, rather than forcing them on users all at once. We speculate that Go may be easier to learn for developers given its garbage collected memory model, which removes some of the burden of memory management from the developers. Could we create a version of Rust with garbage collection as a learning tool?

Reduce the risk of investment. Several of the drawbacks we identified interact in ways that may multiply the perception of risk related to adoption. Much of the cost of adoption occurs up front: the steep learning curve, the relative immaturity of the ecosystem, the slower initial development time, and the inherent challenge of making a large change. Benefits accrue later: improvements in security-minded programming, shorter debugging time and eventually shorter development time overall, and enforced avoidance of key security problems, since Rust is type- and memory-safe. The perceived difficulty of hiring experienced Rust developers, as well as concerns about longevity and future maintenance, may make these future-term benefits seem too uncertain to be worth the risk.

Educators and security advocates who want to incentivize secure programming languages should look for ways to improve this calculus, perhaps by investing in a pipeline of trained Rust developers (reducing learning curve and improving hiring prospects), by developing libraries to contribute to the increasing stability of the ecosystem, or perhaps by developing models and templates for common porting and interoperability challenges. Our participants offer suggestions for action within organizations, such as "starting small" to demonstrate value, and implementing mentoring support for transitioning to Rust. Security advocates could help, by creating and publishing detailed case studies that illuminate benefits and costs of adopting secure tools in real systems, and by creating and supporting mentoring networks for these tools.

Improve the culture around unsafe code. Rust's memory safety-related security benefits come simply by virtue of using the language, but only as long as unsafe blocks are used correctly; the more often and more carelessly they are used, the greater the risk of a security hole. Our participants report that unsafe blocks in Rust code are being used frequently, often with only rudimentary vetting processes; prior studies have come to similar conclusions [3, 10]. While many participants/companies do recognize the risks of unsafe code, we encourage the adoption of more, and more formal, review procedures to more thoroughly mitigate these risks.

Reaching non-enthusiasts. While our participants had a variety of general- and Rust-specific development experience, our sample overrepresents people who show active interest in Rust (as indicated by their adoption efforts and/or membership in a community forum). Future work could explore recruitment strategies to target developers who failed in their adoption efforts and/or lost interest in Rust programming.

9 Related work

Programming language adoption. Chen et al. [5] identified features relevant to a language's adoption success, including institutional support, technology support, and the ability for users to add features. Meyerovich et al. [24] proposed a sociological approach to understanding why some programming languages succeed while others fail. In follow-on work including both project analysis and surveys of developers, they find that open-source libraries, existing code, and prior experience strongly influence developers' selection of languages, while features like performance, reliability, and simple semantics do not [25]. Further, they find that developers tend to prioritize expressivity over correctness. Many of these findings align with our results on the importance of the overall ecosystem to language adoption.

Shrestha et al. [45] studied Stack Overflow questions to understand when and why programmers have difficulty learning a new language, finding that *interference* from previous languages was common, since programmers often attempt to relate a new programming language to ones they know. Our findings suggest that the significant departure from prior experience contributes to Rust's steep learning curve.

Secure tool adoption. Other researchers have investigated factors affecting secure tool adoption by developers. Xiao et al. [50] explored the social factors influencing secure tool adoption, finding that company culture influences adoption and use of security tools through encouragement or discouragement to try new tools and managerial intervention in the security process. In follow-on work, researchers surveyed developers about why they chose (not) to use security tools and found that the biggest predictor of adoption was peers demonstrating the use and benefits of the tool [49]. Haney et al. [17] found that *security advocates* promoting tool adoption must first establish trust by being truthful about risks. These recommendations align with the suggestions our participants offered to Rust advocates.

Other researchers have focused on the adoption of specific tools. Sadowski et al. [44] focused on static analysis tools by building Tricorder which integrates static analysis into developer workflow. They found that developers were generally happy with the results from the static analysis tools and the number of mistakes in the codebase reduced. Christakis et al. [6] explored the factors and features that make a program analyzer appealing to developers by interviewing and surveying developers at Microsoft, finding that the biggest pain-point in using a program analyzer is that the default rules do not match developer wants and developers most want security issues detected.

Real-world Rust usage. Evans et al. [10] studied Rust libraries and applications to uncover how unsafe blocks are used in real-world scenarios and found that less than 30% of Rust libraries contain unsafe blocks, but the most downloaded libraries are more likely than average to use unsafe blocks. Similarly, Astrauskas et al. [3] examine what they call the *Rust hypothesis*: unsafe blocks should be used sparingly, easy to review, and hidden behind a safe abstraction. They find only partial adherence: a large portion of unsafe blocks relate to interoperation, leaving the unsafe blocks publicly accessible and most unsafe blocks are used to call unsafe functions. Qin et al. [41] explored how and why programmers use unsafe blocks, along with the types of security and concurrency bugs found in real Rust programs. They found a number of memory safety issues, all involving the use of unsafe blocks. We explore the use of unsafe blocks, along with mitigation procedures, from the developer's perspective.

Perhaps closest to our work are studies of Rust's usability. Luo et al. [42] developed an educational tool, RustViz, that allows teachers to demonstrate ownership and borrowing by visual example. Mindermann et al. [26] studied the usability of Rust cryptography APIs in a controlled experiment, finding that half of the major cryptography libraries in Rust focus on usability and misuse avoidance. Zeng et al. [51] explored Rust adoption by analyzing Reddit and Hacker News posts relating to Rust, hypothesizing three main barriers to Rust's adoption: tooling which is not promoted by the language developers, difficulty representing complex pointer aliasing patterns, and the high cost of integrating Rust into an existing language ecosystem or toolchain. Our interview and survey study complements this work by asking developers to report on their experiences, positive and negative, with more consistency than can be observed via forum posts but without direct access to specific challenges at the time they occurred. Further, we explore both personal and organizational contexts. Our findings are similarly complementary, identifying both benefits and drawbacks to the tooling and situating the steep learning curve within the eventual benefits.

10 Conclusion

Secure programming languages are designed to alleviate common vulnerabilities that are otherwise difficult to eliminate, such as out-of-bound reads and writes or use-after-free errors. However, these languages cannot provide any security guarantees if they are not adopted. To understand the benefits and hindrances that influence adoption in practice, using Rust as a case-study, we interviewed 16 professional, mostly senior, software engineers who had adopted or tried to adopt Rust on their teams and surveyed 178 members of the Rust developer community. We asked about personal and professional experiences with adopting and using Rust. Participants reported a variety of benefits and drawbacks to adopting Rust, including upfront costs like a steep learning curve and longer-term benefits like shorter development cycles and improved mental models of code security. Participants also discussed reasons their employers were skeptical about Rust adoption, and suggested strategies for championing adoption in the workplace.

11 Acknowledgments

We thank the anonymous reviewers who provided helpful comments on drafts of this paper. This project was supported by NSF grant CNS-1801545.

References

[1] Yasemin Acar, Michael Backes, Sascha Fahl, Simson Garfinkel, Doowon Kim, Michelle L Mazurek, and Christian Stransky. Comparing the usability of cryptographic apis. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 154–171, 2017.

[2] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L Mazurek, and Christian Stransky. You Get Where You're Looking for: The Impact of Information Sources on Code Security. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 289–305, May 2016.

[3] Vytautas Astrauskas, Christoph Matheja, Federico Poli, Peter Müller, and Alexander J Summers. How do programmers use unsafe rust? *PACMPL*, 4(OOPSLA):1–27, 2020.

[4] Yung-Yu Chang, Pavol Zavarsky, Ron Ruhl, and Dale Lindskog. Trend analysis of the cve for software vulnerability management. In *Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 1290–1293. IEEE, 2011.

[5] Yaofei Chen, Rose Dios, Ali Mili, Lan Wu, and Kefei Wang. An empirical study of programming language trends. *IEEE Software*, 22(3):72–79, 2005.

[6] Maria Christakis and Christian Bird. What developers want and need from program analysis: an empirical study. In *Proceedings of the 31st IEEE/ACM international conference on automated software engineering*, pages 332–343, 2016.

[7] Catalin Cimpanu. Microsoft: 70 percent of all security bugs are memory safety issues. https://www.zdnet.com/article/microsoft-70-percent-of-all-security-bugs-are-memory-safety-issues/, 2019.

[8] Juliet Corbin and Anselm Strauss. *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications, 2014.

[9] Ryan Donovan. Why the developers who use Rust love it so much. https://stackoverflow.blog/2020/06/05/why-the-developers-who-use-rust-love-it-so-much/, 2020.

[10] Ana Nora Evans, Bradford Campbell, and Mary Lou Soffa. Is rust used safely by software developers? In *Proceedings of the ACM/IEEE International Conference on Software Engineering*, pages 246–257, 2020.

[11] Deen G Freelon. ReCal: Intercoder reliability calculation as a web service. *International Journal of Internet Science*, 5(1):20–33, 2010.

[12] Alex Gaynor. What science can tell us about C and C++'s security. https://alexgaynor.net/2020/may/27/science-on-memory-unsafety-and-security/, 2020. Presentation at Enigma.

[13] Google. Chrome: 70% of all security bugs are memory safety issues. https://www.chromium.org/Home/chromium-security/memory-safety, 2020.

[14] Google. Go Programming Language. https://golang.org/, 2020.

[15] Jake Goulding. What is Rust and why is it so popular? https://stackoverflow.blog/2020/01/20/what-is-rust-and-why-is-it-so-popular/, 2020.

[16] Greg Guest, Arwen Bunce, and Laura Johnson. How many interviews are enough? an experiment with data saturation and variability. *Field Methods*, 18(1):59–82, 2006.

[17] Julie M. Haney and Wayne G. Lutters. "It's Scary... It's Confusing... It's Dull": How Cybersecurity Advocates Overcome Negative Perceptions of Security. In *Proceedings of the Symposium on Usable Privacy and Security*, 2018.

[18] Andrew F Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89, 2007.

[19] C. A. R. (Tony) Hoare. Null References: The Billion Dollar Mistake. https://www.infoq.com/presentations/Null-References-The-Billion-Dollar-Mistake-Tony-Hoare/, 2009. Presentation at QCon.

[20] IEEE. The Top Programming Languages. https://spectrum.ieee.org/static/interactive-the-top-programming-languages-2019, 2020.

[21] Steve Klabnik and Carol Nichols. The Rust Programming Language Book. https://doc.rust-lang.org/1.9.0/book/README.html, 2020.

[22] Klaus Krippendorff. Reliability in Content Analysis : Some Common Misconceptions and Recommendations. 2015.

[23] Matthew Lombard, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, 28(4):587–604, 2002.

[24] Leo A. Meyerovich and Ariel S. Rabkin. Socio-PLT: Principles for Programming Language Adoption. In *Proceedings of the ACM International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, page 39–54, 2012.

[25] Leo A Meyerovich and Ariel S Rabkin. Empirical analysis of programming language adoption. In *Proceedings of the Conference on Object oriented programming systems languages & applications*, pages 1–18, 2013.

[26] Kai Mindermann, Philipp Keck, and Stefan Wagner. How usable are Rust cryptography APIs? In *Proceedings of the 2018 IEEE International Conference on Software Quality, Reliability and Security (QRS)*, pages 143–154. IEEE, 2018.

[27] Mitre. CVE. https://cve.mitre.org/, 2020.

[28] Mitre. CWE-119: Improper Restriction of Operations within the Bounds of a Memory Buffer. https://cwe.mitre.org/data/definitions/119.html, 2020.

[29] Mitre. CWE-125: Out-of-bounds Read. https://cwe.mitre.org/data/definitions/125.html, 2020.

[30] Mitre. CWE-416: Use After Free. https://cwe.mitre.org/data/definitions/416.html, 2020.

[31] Mitre. CWE-476: NULL Pointer Dereference. https://cwe.mitre.org/data/definitions/476.html, 2020.

[32] Mitre. CWE-787: Out-of-bounds Write. https://cwe.mitre.org/data/definitions/787.html, 2020.

[33] Mozilla. Rust Programming Language. https://www.rust-lang.org/, 2020.

[34] Mozilla. The Rust programming language. https://developer.mozilla.org/en-US/docs/Mozilla/Rust, 2020.

[35] Mozilla. Rust Programming Language Production. https://www.rust-lang.org/production, 2020.

[36] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, Emanuel von Zezschwitz, and Matthew Smith. "If You Want, I Can Store the Encrypted Password": A Password-Storage Field Study with Freelance Developers. In *Proceedings of the Conference on Human Factors in Computing Systems*, pages 140:1–140:12, 2019.

[37] NIST. CWE Over Time. https://nvd.nist.gov/general/visualizations/vulnerability-visualizations/cwe-over-time, 2020.

[38] NIST. National Vulnerability Database. https://nvd.nist.gov/general, 2020.

[39] Jeffrey M Perkel. Why scientists are turning to Rust? *Nature*, 588:186–186, 2020.

[40] Rob Pike. Go at Google: Language design in the service of software engineering. https://talks.golang.org/2012/splash.article, 2020.

[41] Boqin Qin, Yilun Chen, Zeming Yu, Linhai Song, and Yiying Zhang. Understanding Memory and Thread Safety Practices and Issues in Real-World Rust Programs. In *Proceedings of the Conference on Programming Language Design and Implementation*, page 763–779, 2020.

[42] Vishnu Reddy, Marcelo Almeida, Yingying Zhu, Ke Du, Cyrus Omar, et al. RustViz: Interactively Visualizing Ownership and Borrowing. *arXiv preprint arXiv:2011.09012*, 2020.

[43] Andrew Ruef, Michael Hicks, James Parker, Dave Levin, Michelle L Mazurek, and Piotr Mardziel. Build it, break it, fix it: Contesting secure development. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 690–703, 2016.

[44] Caitlin Sadowski, Jeffrey Van Gogh, Ciera Jaspan, Emma Soderberg, and Collin Winter. Tricorder: Building a Program Analysis Ecosystem. In *Proceedings of the 2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, volume 1, pages 598–608, 2015.

[45] Nischal Shrestha, Colton Botta, Titus Barik, and Chris Parnin. Here We Go Again: Why Is It Difficult for Developers to Learn Another Programming Language? In *Proceedings of the ACM/IEEE International Conference on Software Engineering*, 2020.

[46] StackOverflow. Developer Survey Results. https://insights.stackoverflow.com/survey/2020#technology-most-loved-dreaded-and-wanted-languages-loved, 2020.

[47] Laszlo Szekeres, Mathias Payer, Tao Wei, and Dawn Song. Sok: Eternal war in memory. In *Proceedings of the IEEE Symposium on Security and Privacy*, pages 48–62, 2013.

[48] Jim Witschey, Shundan Xiao, and Emerson Murphy-Hill. Technical and personal factors influencing developers' adoption of security tools. In *Proceedings of the ACM Workshop on Security Information Workers*, pages 23–26, 2014.

[49] Jim Witschey, Olga Zielinska, Allaire Welk, Emerson Murphy-Hill, Chris Mayhorn, and Thomas Zimmermann. Quantifying developers' adoption of security tools. In *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*, pages 260–271, 2015.

[50] Shundan Xiao, Jim Witschey, and Emerson Murphy-Hill. Social influences on secure development tool adoption: why security tools spread. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1095–1106, 2014.

[51] Anna Zeng and Will Crichton. Identifying barriers to adoption for Rust through online discourse. *arXiv preprint arXiv:1901.01001*, 2019.

A Interview protocol

Most interviews were conducted by one interviewer; some interviews were assisted by a second interviewer. The second interviewer took notes and asked some additional and follow-up questions. Each interview was audio recorded, with permission.

Rust Background

- How did you initially learn about Rust?
- Why did you/your team/your company decide to adopt Rust?
- Was it hard to convince the necessary people/groups (bosses, team members, others?) at your company to use Rust?
  - What concerns did they have?
  - What were they excited about?
- Have any attitudes/policies of (team members, management) changed since you attempted this project in Rust?
  - How so?
- Can you please describe at a high level the project you/your team/your company are/is working on in Rust?
  - Is the project currently ongoing?
    - If yes, would you (briefly) characterize it as going well? Why (not)?
    - If no, did you finish it?
      - If no, why do you think you weren't able to complete the project?
      - If yes, do you consider the outcome a success? Why (not)?
  - Why did you pick this project to write in Rust?
- Can you tell me more about what happened when you tried to adopt Rust?
  - How long did it take you/did you spend trying/do you think you'll need to complete this project in Rust?
  - What went particularly well when adopting Rust at your company/on your team?
  - What went particularly poorly when adopting Rust at your company/on your team?
  - Did you receive positive feedback from adopting Rust?
    - From whom?
    - What were they happy about?
  - Did you receive negative feedback from adopting Rust?
    - From whom?
    - What were they happy about?
  - What would you do differently if you were to attempt another project in Rust?
    - Would you even try again at all?
- What would you tell someone in your position at a different company that is also thinking about adopting Rust?

Experiences with Rust (Only ask if they are programmer/familiar with coding)

- Have you ever felt like there was a programming task or something you wanted to program in Rust but could not get it to work?
  - What was it?
  - What did you try in order to debug/fix this problem?
- Can you tell me how Rust specific things affect your ability to fit a problem specification into a solution in Rust?
  - Ownership?
  - Lifetimes?
- Can you tell me more about your process of going from a problem specification to a solution in Rust?

- Do you find it difficult to find a solution to a programming problem in Rust?
  - Why?
- How easy is it for you to find solutions to any problems or errors you encounter while programming in Rust?
- What features do you like most about the Rust programming language?
  - Libraries/APIs?
  - Online community?
- What features do you like least about the Rust programming language?
  - Libraries/APIs?
  - Online community?
- In your opinion, what are the biggest strengths of Rust?
  - Libraries/APIs?
  - Online community?
- In your opinion, what are the biggest weaknesses of Rust?
  - Libraries/APIs?
  - Online community?
- Have you ever used unsafe blocks in this project?
  - Why did you use them?
  - What solutions did you try before using the unsafe blocks?
- Does this project code interoperate with any other code from another language?
  - What was hard about getting the code to interoperate?
  - What was easy about getting the code to interoperate?
- Did you/the team port code from another programming language to Rust for this project?
  - What language did you port from?
  - What was hard about porting your code to Rust?
  - What was easy about porting your code to Rust?
  - Did you feel like it was easier to write this code in the original language or Rust?

B Survey

Technical Background

1. How long have you been programming? **[Less than a year, 1 - 5 years, 5 - 10 years, More than 10 years]**
2. Are you currently employed in a software engineering field? **[Yes, Maybe, No]**
3. Which of the following currently describe(s) what you do for work? (Check all that apply) (*Only show if they answered yes or maybe to question 2*) **[Operating systems programming, Embedded systems programming, Firmware development, Web development, Network programming, Databases programming, Game development, Data science, DevOps, Desktop/GUI applications development, Library development, Mobile application development, CS/Technical research, CS/Technical education, Other [text box]]**
4. Approximately how many employees work for your employer? (*Only show if they answered yes or maybe to question 2*) **[1 - 100, 100 - 999, 1000 or more]**
5. Which of the following programming languages have you been using for the last year (in a substantive manner), and/or expect to use in the near term? (Check all that apply) **[Python, C++, Java, C, C#, PHP, R, Javascript, Swift, Go, Haskell, Other (Please comma separate if more than 1) [text box]]**
6. Please rate your level of comfort and experience using the following programming languages.

   1 - I have never used the programming language.

   2 - I have used the programming language sparingly (e.g. modifications to others' programs or small toy programs)

   3 - I have written a few thousand lines of code in the programming language.

   4 - I am comfortable writing in it.

   5 - I have programmed in this language a lot and know it very well.

   [Python, C++, Java, C, C#, PHP, R, Javascript, Swift, Go, Haskell]
7. To what extent have you used the Rust programming language? **[I have used Rust for hobby projects, I have used Rust in a class, I have maintained a body of Rust code, I have been paid to write Rust code, I have never used Rust]**
8. What were the main reason(s) that you decided to learn Rust? (Check all that apply) **[Rust was assigned for a class, Rust was assigned for a paid job, Curiosity**

about Rust, Rust was suggested by a friend, To learn a marketable job skill, Other [text box]]

9. How long have you been programming in Rust? (Total time which you have actively spent working on Rust projects) **[Less than a year, 1- 2 years, 2 - 5 years, More than 5 years]**

10. How many lines of code (LOC) do you estimate you have written in Rust? **[0 - 1000 LOC, 1000 - 10k LOC, 10k - 50k LOC, 50k - 100k LOC, More than 100k LOC]**

11. Which best describes your current use of Rust? **[I am currently using Rust for projects., I have used Rust in the past for projects, but I am not using it currently., I am not currently using Rust for projects.]**

12. If it were up to you to choose, how would you feel about using Rust for future projects? **[I would definitely want to use Rust in the future., I probably want to use Rust in the future., I probably do not want to use Rust in the future., I definitely do not want to use Rust in the future.]**

13. Notwithstanding your general interest in using Rust in the future, for which of the following tasks/applications/projects would you not choose Rust? (Check all that apply) **[GUI applications, Web applications, Mobile applications, Writing compiler code, Writing graphics code, Writing testing code, Other [text box]]**

Learning and Using Rust

1. Which of the following describes the primary way(s) you learned Rust? (Check all that apply) **[Followed a Rust tutorial, Worked through "The Rust Programming Language" on-line text, Asked questions about Rust through on-line forums, Asked questions about Rust to coworkers/group-mates/friends, Studied Rust in a class, Attended a Rust workshop/bootcamp, Wrote a small Rust program from scratch, Ported some existing code to Rust, Other [text box]]**

2. How easy or difficult did you find Rust to learn? **[Very difficult, Slightly difficult, Neither difficult nor easy, Slightly easy, Very easy]**

3. How long after learning and using Rust did it take before you could quickly and easily write a program that compiled and ran (without frequently resorting to the use of unsafe blocks)? **[Less than 1 week, 1 week - 1 month, 1 month - 6 months, 6 months - 1 year, More than 1 year, I am not yet able to quickly and easily write a program that compiles and runs.]**

4. Approximately how long did it take for you to feel comfortable in Rust writing:
   - A small program (Less than 10,000 lines of code) **[1 week, 1 week - 1 month, 1 month - 6 months, 6 months - 1 year, More than 1 year, N/A]**
   - A large program (More than 10,000 lines of code) **[1 week, 1 week - 1 month, 1 month - 6 months, 6 months - 1 year, More than 1 year, N/A]**
   - Library code [1 week, 1 week - 1 month, 1 month - 6 months, 6 months - 1 year, More than 1 year, N/A]
   - An application **[1 week, 1 week - 1 month, 1 month - 6 months, 6 months - 1 year, More than 1 year, N/A]**

5. How would you rate the quality of available Rust documentation? **[Very poor, Poor, Average, Good, Very good, I don't know]**

6. How would you rate the quality of advice from the Rust online community (For example: reddit, Stack Overflow, etc)? **[Very poor, Poor, Average, Good, Very good, I don't know]**

7. To what extent do you agree with the following statement: When I encounter a problem or error while working in Rust, I can easily find a solution to my problem? **[Strongly disagree, Disagree, Neither agree nor disagree, Agree, Strongly agree, I don't know]**

8. Which of the following make(s) the process of finding a solution to your problems or errors easy? (Check all that apply) (*Only show if the answer to 7 is agree or strongly agree*) **[Availability of examples in official documentation, Availability of examples on Stack Overflow, Availability of examples on other online tutorials, Availability of knowledgable teammate/friend, Availability of descriptive compiler/error messages, Strong understanding of the language, Other [text box]]**

9. Which of the following make(s) the process of finding a solution to your problems or errors difficult? (Check all that apply) (*Only show if the answer to 7 is disagree or strongly disagree*) **[Lack of examples in official documentation, Lack of examples on Stack Overflow, Lack of examples on other online tutorials, Lack of knowledgable teammate/friend, Lack of descriptive compiler/error messages, Lack of strong understanding of the language, Other [text box]]**

Using Rust for Work

1. Are you, personally, currently writing Rust code for work? **[Yes, No]**

2. Which of the following most accurately describes how you are writing Rust code for work? (*Only show if the answer to 1 is yes*) **[I am writing Rust code as part of a company or large organization., I am writing Rust code as part of a freelance assignment.]**

3. Have you or anyone on your team tried to get Rust adopted at your employer? (*Only show if the answer to 1 is no*) **[Yes, No]**

4. What were the major reasons your employer stated for deciding against using Rust? (Check all that apply) (*Only show if the answer to 3 is yes*) **[Insufficient security, Inadequate performance, Lack of interoperability with existing codebase, Difficulty of maintainability, Lack of compatibility with development ecosystem, Difficulty of learning the language, Potential reduction in productivity of developers, Unfamiliarity with the language, Inability to hire Rust developers, Lack of trust in Rust toolbase, Concern about the long-term development and support of the language, Time pressure to deliver a product, Not wanting another new language at the company, Other [text box]]**

5. Did anyone at your employer/on your team have apprehensions about using Rust? (*Only show if the answer to 2 is I am writing code as part of a company or large organization*) **[Yes, No]**

6. What were the major apprehensions of your employer/teammate(s) about using Rust? (Check all that apply) (*Only show if the answer to 5 is yes*) **[Insufficient security, Inadequate performance, Lack of interoperability with existing codebase, Difficulty of maintainability, Lack of compatibility with development ecosystem, Difficulty of learning the language, Potential reduction in productivity of developers, Unfamiliarity with the language, Inability to hire Rust developers, Lack of trust in Rust toolbase, Concern about the long-term development and support of the language, Time pressure to deliver a product, Not wanting another new language at the company, Other [text box]]**

7. Other than Rust, what language(s) do you primarily use at your employer (in terms of largest number of projects and/or lines of code)? (Check all that apply) (*Only show if the answer to 2 is I am writing code as part of a company or large organization* **[Python, C++, Java, C, C#, PHP, R, Javascript, Swift, Go, Haskell, Other [text box]]**

8. What language(s) do you primarily use at your employer (in terms of largest number of projects and/or lines of code)? (Check all that apply) (*Only show if the answer to 1 is no and 3 is yes* **[Python, C++, Java, C, C#, PHP, R, Javascript, Swift, Go, Haskell, Other [text box]]**

9. What are the primary conditions under which you have developed using Rust at your employer? (*Only show if the answer to 2 is I am writing code as part of a company or large organization*) **[Developing alone, Developing in teams of 2 - 5 people, Developing in teams of more than 5 people]**

10. Approximately how much legacy code at your employer was written in another language? (*Only show if the answer to 2 is I am writing code as part of a company or large organization*) **[Less than 100,000 lines of code, 100,000 - 1,000,000 lines of code, 1,000,001 - 500,000,000 lines of code, 500,000,001 - 1,000,000,000 lines of code, More than 1,000,000,000 lines of code]**

11. Which best describes your employer's future use of Rust after the completion of current project(s), if any? (*Only show if the answer to 2 is I am writing code as part of a company or large organization*) **[I am certain that my employer will use Rust again in the future., I think my employer will use Rust again in the future., I do not think my employer will use Rust again in the future., am certain my employer will not use Rust again in the future.]**

12. What one piece of advice would you give to someone who is just starting out in writing Rust at an employer similar to yours? (*Only show if the answer to 2 is I am writing code as part of a company or large organization*) **[text box]**

13. What one piece of advice would you give to someone who is trying to get Rust adopted at an employer similar to yours? (*Only show if the answer to 3 is yes*) **[text box]**

Comparing Rust to Other Languages

1. The next set of questions will ask you to compare your opinions about and experiences with Rust to those of another language. This language should be among those you are most comfortable programming in; it can be your favorite, or perhaps the one you are using most right now. Please choose it from the list below. **[Python, C++, Java, C, C#, PHP, R, Javascript, Swift, Go, Haskell, N/A (Rust is the only language I program in), Other [text box]]**

2. How would you rate the **quality of Rust debugging tools** compared to [*chosen language*]? **[Very poor, Poor, Average, Good, Very good]**

3. How would you rate the **quality of Rust testing tools** compared to [*chosen language*]? **[Very poor, Poor, Average, Good, Very good]**

4. How would you rate the **quality of Rust compiler and run-time error messages** compared to [*chosen language*]? **[Very poor, Poor, Average, Good, Very good]**

5. To what extent do you agree with the following statement: I can **more quickly design and fully implement code in Rust** (well-tested, few if any bugs) than in [*chosen language*]? **[Strongly disagree, Slightly disagree, Neither agree nor disagree, Slightly agree, Strongly agree]**

6. To what extent do you agree with the following statement: I find it **more difficult to prototype in Rust** (i.e., get the basic working, but there may be bugs and missing corner cases) than in [*chosen language*]? [Strongly disagree, Slightly disagree, Neither agree nor disagree, Slightly agree, Strongly agree]

7. To what extent do you agree with the following statement: I spend **more time getting my Rust code to compile** than code in [*chosen language*]? [Strongly disagree, Slightly disagree, Neither agree nor disagree, Slightly agree, Strongly agree]

8. To what extent do you agree with the following statement: Once I get it to compile. I spend **less time debugging my Rust code** than code in [*chosen language*]? **[Strongly disagree, Slightly disagree, Neither agree nor disagree, Slightly agree, Strongly agree]**

9. To what extent do you agree with the following statement: **Rust code is more difficult to maintain** than code in [*chosen language*]? **[Strongly disagree, Slightly disagree, Neither agree nor disagree, Slightly agree, Strongly agree]**

10. To what extent do you agree with the following statement: **Rust makes me more confident that my production code is bug-free** than programming in [*chosen language*]? **[Strongly disagree, Slightly disagree, Neither agree nor disagree, Slightly agree, Strongly agree]**

11. To what extent do you agree with the following statement: **Rust reduces the amount of time from the start of a project to shipping the project** compared to [*chosen language*]? **[Strongly disagree, Slightly disagree, Neither agree nor disagree, Slightly agree, Strongly agree]**

Rust Language/Ecosystem

1. Recall recent experiences developing with Rust. What are some things about Rust - both language and ecosystem - that you **liked**? (Check all that apply) **[Traits, Slices, Enums, Memory safety, Concurrency safety, Immutability by default, Pattern matching, No null pointers, Closures, Generics, Ownership, Lifetimes, Performance, Crates ecosystem, Lack of garbage collection, Other [text box]]**

2. Recall recent experiences developing with Rust. What are some things about Rust - both language and ecosystem - that you **disliked**? (Check all that apply) **[Dependency bloat, Lack of available libraries, Long build time, Code size, Prototyping in Rust, Missing features (Please elaborate below) [text box], Other [text box]]**

3. Which of the following describes your use of unsafe blocks/code while programming in Rust? (Check all that apply) **[I have used unsafe code for foreign function interface (FFI) code., I have used unsafe code to enhance the performance of my code., I have used unsafe code for kernel-level/low-level interaction., I have used unsafe code for hardware interaction., I have used unsafe code to allow for memory management., I have used unsafe code in another way. (Please elaborate below) [text box], I have never used unsafe code.]**

4. Does your employer/team/do you have a system for the review and use of unsafe blocks? (*Only show if the answer to 3 is not I have never used unsafe blocks*) **[Yes [text box], No]**

5. To what extent has Rust positively affected how you program in [*chosen language*]? **[No effect at all, Minor effect, Some effect, Moderate effect, Major effect]**

6. How has Rust affected how you work in [*chosen language*]? (Check all that apply) (*Only show if the answer to 5 is not no effect at all*) **[Made me think about ownership, Made me think about aliasing, Made me think about memory lifetimes, Made me think about data structure organization, Made me think about the use of generics, Made me think about the use of immutability, Made me think about iteration patterns within my code, Other [text box]]**

Porting and Interoperating with Legacy Code

1. Have you ever tried to port code from another language into Rust? **[Yes, No]**

2. What language(s) have you ported from? (Check all that apply) (*Only show if they answered yes or maybe*

to question 2) **[Python, C++, Java, C, C#, PHP, R, Javascript, Swift, Go, Haskell, Other [text box]]**

3. How easy or difficult was it to port code from[*chosen language*] to Rust? **[Extremely easy, Somewhat easy, Neither easy nor difficult, Somewhat difficult, Extremely difficult]**

4. Have you ever written Rust code intended to interoperate with code in another programming language? **[Yes, No]**

5. What language(s) have you tried to interoperate with? (Check all that apply) (*Only show if they answered yes or maybe to question 4*) **[C, C++, Other [text box]]**

6. How easy or difficult was it to achieve the interoperation between [*chosen language*] and Rust? **[Extremely easy, Somewhat easy, Neither easy nor difficult, Somewhat difficult, Extremely difficult]**

Background of Participants

1. Please select your gender: **[Male, Female, Non-binary, Other [text box], Prefer not to answer]**

2. Please select your age: **[18 - 29, 30 - 39, 40 - 49, 50 - 59, 60 - 69, Over 70, Prefer not to answer]**

3. Please select your highest completed education level: **[Some high school, High school diploma/GED, Some college, Bachelor's degree, Master's degree, PhD]**

C Demographic tables

ID	Title	Company Size (# employees)
I1	aid in Rust adoption	$\geq$ 1000
I2	group director	100 - 999
I3	software engineer	$\geq$ 1000
I4	software engineer	$\geq$ 1000
I5	senior engineer	100-999
I6	principal software engineer	$\geq$ 1000
I7	system engineer	$\geq$ 1000
I8	software engineer	$\geq$ 1000
I9	co-founder and CTO	$<$ 100
I10	instrumentation engineer	100 - 999
I11	engineering manager	$\geq$ 1000
I12	research software engineer	100 - 999
I13	research software engineer	100 - 999
I14	software engineer	$\geq$ 1000
I15	principal engineer	$<$ 100
I16	software engineer	$\geq$ 1000

Table 1: Interviewee demographics.

Area	# of participants (%)
Web development	73 (54%)
Library development	51 (38%)
Network programming	44 (32%)
DevOps	33 (24%)
Databases programming	28 (21%)
Data science	27 (20%)
Embedded systems programming	27 (20%)
Desktop/GUI apps development	26 (19%)
OS programming	25 (18%)
Other	24 (18%)
Mobile application development	19 (14%)
Firmware development	16 (12%)
CS/Technical research	14 (10%)
Game development	9 (7%)
CS/Technical education	7 (5%)

Table 2: Survey participants who worked in each area of software development. Multiple selection was allowed.

An Analysis of the Role of Situated Learning in Starting a Security Culture in a Software Company

Anwesh Tuladhar Daniel Lende Jay Ligatti Xinming Ou
University of South Florida, Tampa, FL, USA
Email: {atuladhar, dlende, ligatti, xou} @usf.edu

Abstract

We conducted an ethnographic study of a software development company to explore if and how a development team adopts security practices into the development lifecycle. A PhD student in computer science with prior training in qualitative research methods was embedded in the company for eight months. The researcher joined the company as a software engineer and participated in all development activities as a new hire would, while also making observations on the development practices. During the fieldwork, we observed a positive shift in the development team's practices regarding secure development. Our analysis of data indicates that the shift can be attributed to enabling all software engineers to see how security knowledge could be applied to the specific software products they worked on. We also observed that by working with other developers to apply security knowledge under the concrete context where the software products were built, developers who possessed security expertise and wanted to push for more secure development practices (security advocates) could be effective in achieving this goal. Our data point to an interactive learning process where software engineers in a development team acquire knowledge, apply it in practice, and contribute to the team, leading to the creation of a set of preferred practices, or "culture" of the team. This learning process can be understood through the lens of the situated learning framework, where it is recognized that knowledge transfer happens within a community of practice, and applying the knowledge is the key in individuals (software engineers) acquiring it and the community (development team) embodying such knowledge in its practice. Our data show that enabling a situated learning environment for security gives rise to security-aware software engineers. We discuss the roles of management and security advocates in driving the learning process to start a security culture in a software company.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

1 Introduction

A wide range of research has addressed how to best establish secure development practices within a software development team/company. The standard approach is to use a secure development model to formulate a suitable secure software development lifecycle (S-SDLC) [10,19,36,43]. Despite the success of S-SDLC in its originating company [19], successful establishment of secure development practices has remained more difficult for the software industry at large. In general, some companies are unwilling to place code security on a level playing field with business considerations such as time-to-market of the product. There are also companies that make an effort to deliver secure code through the adoption of secure development life cycles but have not been able to do so effectively. Reasons for such failures have been posited as lack of security knowledge in developers, lack of available resources, lack of usable security, improper use of security APIs, and so on [1, 12, 33, 35]. Use of security tools is often suggested as a way to help alleviate such problems by catching developer mistakes before they land in the product. However the adoption of security tools into development itself can remain an issue [34]. Some studies allude to the notion of security mindset or security culture within the company as the influential factor in driving these secure development practices [5, 17, 42]. But what is a security culture? What are the benefits it provides and how can a company start to develop such a culture?

We conducted an extensive ethnographic study in a software company in order to understand how and why secure coding practices are (or are not) integrated into software development processes. We embedded a computer science PhD student with training in qualitative methods as part of the

company's software development team. Approximately two months prior to the researcher joining, the company had initiated the process of implementing a secure development lifecycle, with upper-management declaring that security should receive greater consideration. This new emphasis on security provided us an opportune moment to conduct research into whether and how such push for security may result in concrete positive changes in the development processes. Our research was helped by the company empowering a recently hired software engineer who was also trained in security, to push for secure development practices within the development team. The main contributions of our work are as follows.

1. We identify an important factor in establishing secure development practices in a software company – the role of situated learning [20] that forms an integral part of software engineers' work. Rather than assuming structured processes on their own can solve security problems, we examine the context for learning about security within the development environment and analyze how it shapes the workflows followed by individual software engineers. Our analysis of data shows that what was driving the positive shift in the development team's security awareness can be explained by the learning dynamics existent therein, in particular the software engineers being able to identify the applicability of the security knowledge within the context of the everyday work they perform. The situated learning dynamics could drive the team into a set of agreed-upon knowledge and the associated practices, becoming the "preferred practices" for dealing with specific security concerns. We hence identify a way to start a secure coding culture in a development team.

2. Our data also indicate that the presence of a security expert working within the development team is instrumental in driving the situated learning cycle for security. It appears that when such security experts are part of the development team, and their actions foster the learning process, the adoption of secure coding practices become more readily accepted by the team. In particular, we find it important that security knowledge be offered within the context of the team's concrete work.

2 Fieldwork

Our fieldwork was conducted at a software development company headquartered in the United States with offices throughout the world. The researcher was embedded in a development team responsible for two security-related products developed by the company, referred to as P1 and P2 in this paper. Due to the ongoing COVID-19 pandemic, the mode of work varied between work-from-home and on-premise. The company followed local government guidelines; mandated work-from-home when stay-at-home order was in effect, and provided the flexibility of either work-from-home or on-premise otherwise. For on-premise work, the company followed all safety precautions by reorganizing the office setup to socially distance cubicles and providing masks and hand sanitizers. All meetings were held through video conferencing even when on premise. The advantages of being on premise were ease of access to the test environment and ability to start impromptu discussions and meetings when necessary.

2.1 The Development Team

The main participants were five software engineers (SWEs) in the development team, two network engineers, two support engineers, two sales/customer relations representatives, one quality assurance engineer (QAE), one graphic designer, and one vice president (VP), who also oversaw the management of the two products. All SWEs had at least 1.5 years' experience within the company, with two having more than five years' experience. The QAE joined during the fieldwork.

2.2 Research Methodology

We employed the qualitative research method of participant observation [7, 29]. In this method, the researcher spends an extended amount of time in the field taking part in day to day activities and practices. This method allows researchers to obtain in-depth understanding of practices such as software development that take place over long periods of time.

The participant observer in this research was a computer science PhD student with prior training in qualitative research methods as well as ample industry experience. This experience allowed the researcher to integrate quickly into the daily work. The researcher spent three days per week at the company for a duration of eight months. Although the researcher's background was in security, the researcher was not limited to security-related tasks and participated in all activities a regular SWE at the company would, such as sprint planning, scrum meetings, bug fixes, feature design/implementation/testing, and code reviews.

Our data analysis utilized the general inductive approach [30], augmented by specific techniques for qualitative data analysis such as analytic notes and comparative analysis [6]. The researcher maintained descriptive field notes on daily activities and interactions. After three months of data collection, the research team met weekly to reflect on the observations made so far. The team went over the events of the past week and discussed the events concerning software development practices, security practices, and the relevant interactions. For the security incidents encountered, we separately kept track of the process of identification, technical details of the issue, and the progress made towards mitigating them. The researcher coded the raw field data based on the patterns and themes that emerged during these discussions. Any unanswered questions and/or missing information during

these discussions then guided the future observations in the field. The weekly iteration of data collections followed by in-depth discussions led to the refinement of the emerging categories used for coding (see Appendix for the final set of codes used in our research). Then, as broader themes were conceptualized, the researcher started to write analytic notes summarizing each theme and documenting ideas and analysis of each along with supporting data from exploration of code, tickets, and other relevant sources in addition to the raw fieldwork data. Multiple themes emerged and evolved throughout this process. After the end of the fieldwork, the research team continued further analysis of data through extensive discussions to draw out the major implications of our observations to secure software development practices.

The study was reviewed and approved by the Institutional Review Board (IRB). The researcher explained the study goals to participants and obtained verbal informed consent from them. Field notes were anonymized, as well as discussions during weekly research meetings.

3 Software Development Processes and Challenges Facing Secure Development

Approximately two months prior to the researcher joining the development team, management instructed the team to employ secure software development lifecycle (S-SDLC). This provided an invaluable opportunity for the research team to examine whether and how secure development practices can take hold in a software development team when there is buy-in from the top. In this section, we describe our observations of the company's overall software development processes and challenges facing secure development throughout our fieldwork. In section 4 we focus our discussion on observations and analysis of the shift in the development processes as a result of the management push for S-SDLC.

The company adopted a sprint-based agile development model. An issue-tracking tool was used for planning and tracking the development progress throughout a sprint. We describe this process below.

3.1 Sprint Planning

Everyone in the team was free to create a ticket for any work that was not already tracked and would be added to the *backlog* queue. However, only the lead SWEs could approve the ticket for development. In addition, the VP and customer facing specialists could add feature tickets based on company vision and customer requests/feedback. Each week the VP and the lead sales representative, along with the lead SWEs had a *prioritization meeting* where the new tickets were discussed, approved/rejected for future development, with the approved tickets ranked based on priority. For each sprint, the SWEs and QAE conducted a *sprint planning* meeting where the highest priority tickets from the backlog were discussed and assigned *story points* representing the estimated complexity/amount of work. Story points for each ticket were agreed upon by the whole team using SCRUM poker [41], where each SWE and QAE anonymously assigned story points based on their understanding of the required work. When the assigned scores varied widely, a discussion was held to allow each SWE/QAE to explain the reasoning for their scores and SCRUM poker was re-done, until everyone converged on a common score. A total of 60-70 story points were targeted for each sprint, which allowed for a small number of additional high-priority tickets or unforeseen issues to be included in the sprint at a later time.

3.2 Development Workflow

A short 20-30 minute scrum meeting was held every morning to provide brief updates on the progress from the previous day, any issues/roadblocks encountered, and goals for the current day. The meeting was led by the lead SWE and included all SWEs, the QAE, the graphic designer, and the VP. Additional meetings were called by individuals as required to discuss ticket requirements, design issues, knowledge transfer for codebase, or testing strategies. We next discuss the stages of development and the challenges facing secure development in the context of each stage.

3.2.1 Design

A high-level design discussion was held during the sprint planning. For simple tickets, the SWE assigned took the responsibility of finalizing the design. For more complex tickets, discussions were held with the appropriate team members. In some cases, a wiki page with suggested alternatives was requested before such discussions.

Including security as a part of the design consideration presented the following challenges.

Challenges regarding security knowledge of SWEs. During the design stage the main focus was to achieve functional correctness and performance considerations when applicable. When the features dealt with sensitive information, security became a necessity. Yet, secure design practices were not always the focus and instead assumed protection through "security functions" such as authentication and authorization. For example, one SWE when asked if the input attributes should be validated:

> *"I'm not sure I'd worry too much about that. This form is authorized for admin only, so customers won't be changing this attribute themselves. Trying to validate that they've provided a valid <redacted> attribute feels kind of complicated..."*

Secure design must determine relevant threats (through threat modeling) and consider all aspects of the software, but it is a tall order to require all SWEs to have such knowledge and skills.

Challenges in understanding contextual knowledge by a security expert. One of the software engineers in the team had a significant background in security (called SecSWE hereafter). He was able to identify the relevant security risks and propose secure design solutions. However, although he had been working with the team for more than 1.5 years, the knowledge of the minutiae of how the product operated was lacking and the proposed solutions were not always directly applicable for the product at hand. Since there was no one-size-fits-all solution to security problems, secure design required in-depth knowledge of both security and the product.

3.2.2 Implementation

SWEs picked up tickets from the sprint plan to implement. Again, the first priority of SWEs was to implement the functional requirements of the tickets. We observed the following secure development challenges in this stage.

Challenges regarding security knowledge in SWEs. Even with security considerations in the design phase, the actual implementation of code could expose vulnerabilities if the SWE was not capable of defensive programming and unaware of secure development practices. The main challenge is that the SWE needs to be able to identify the potential security risks in the code that he/she is writing. Other factors such as reliance on frameworks, or incorrect use of frameworks or APIs can also lead to insecure implementation. Such lack of knowledge in SWEs cannot be simply compensated by the presence of a security expert in the team. Usually in such cases the identification of security issues shifts further down the development process during code analysis or security testing and presents additional challenges – the issue might be missed altogether, fixing the security could require significant code changes or even design changes, or there might not be enough time in the sprint to fix the issues which might lead to the ticket being excluded from the sprint.

Challenges in applying security knowledge in practice. In certain cases SWEs were able to identify potential security risks and also had the necessary knowledge to resolve them, but chose not to do so. We concluded this based on our data where we observed that during discussions regarding security issues, often times some SWEs were able to propose the solutions, but did not apply them in practice. This could be attributed to multiple reasons such as lack of time, reliance on security functions (such as authentication and authorization), or security of code considered "invisible" to the customer compared to the feature itself.

3.2.3 Continuous Integrations/Code Analysis

Once the implementation was complete, the source code was pushed to the remote repository where automated builds were carried out by the continuous integration (CI) pipelines. These pipelines executed unit/integration tests and code analysis tools such as SonarQube [37] (later Black Duck [44]) on the feature branch.

Challenges on fully utilizing available tools. We found that the available tools were not utilized to their full capabilities. SonarQube was not maintained and mostly only relied upon for simply lint and code quality checks. The team was asked to set up Black Duck, a tool that analyzes the use of third-party open-source libraries in the codebase and provides information on licenses and known security vulnerabilities, into the CI pipeline as a part of the secure development effort. During discussions on the initial results of the scan, one SWE remarked: *"We use quite a few out-dated packages. I would be surprised if this tool didn't report any issues."* Black Duck was setup on management's request, and the scans were initially enabled by default in the CI pipeline. But the resulting build failures in the Black Duck stage prevented SWEs from merging in code. The tool was then disabled by default and a separate ticket was filed to track and address the vulnerabilities discovered.

3.2.4 Developer Testing

Before a ticket was assigned for code review, SWEs made sure that all automated tests were passing, deployed the updated product on a test environment, and performed their own testing. These tests were usually targeted towards functionality rather than security. Once functionality was verified, the ticket was updated with the steps to replicate the test plan and assigned for code review with the creation of a pull request.

3.2.5 Code Review

Two SWEs were assigned for code review. Usually one SWE provided thorough review while the other would just sanity-check the code. Depending on the complexity of the feature, the reviewers may perform quick functionality tests on top of going through the code changes. Any missing pieces, mistakes, inconsistency or departure from existing best practices in the coding pattern were set up as tasks to be addressed before the ticket was marked as "done." We observed the following challenges in this stage.

Challenges in consistently performing code review. Occasionally, when the changes were required urgently, code review was essentially skipped with the SWE just describing the changes made to others and asking if anyone objected to the approach. This could lead to potentially identifiable issues propagating to the production code base.

Challenges in thoroughly performing security review. Although SWEs provided good feedback during code review, the suggested changes were based on internal best practices and patterns followed in other similar modules in the product. However, a thorough security code review requires more in-depth security knowledge and experience which was lacking in the SWEs. SecSWE however was able to provide specific

security-related feedback. A potential API misuse of a crypto-library (bouncycastle [39]) was identified by SecSWE during code review. While addressing this comment, it was discovered that the API misuse could have caused memory leaks leading to out of memory conditions.

3.2.6 Post Development Testing

The QAE, who was hired during our fieldwork, prepared thorough test plans for each ticket in the sprint and carried them out on the test build. We observed the following challenges at this stage.

Challenges in acquiring contextual knowledge by QAE. Although the QAE had years of prior experience, he was new to the team and the products, and hence required assistance to set up test environments and understand the specifics of the product before he could create strategic test plans. The QAE also had a security background and showed interest to learn and practice security-oriented testing along with SecSWE. But he expressed lack of time and in-depth knowledge about the product as reasons not to do so at that time.

3.3 Product Release

At the end of a sprint, a build of the product including all implementations in the sprint was deployed within the company for up to a week; then release notes were written and the product was released. The customers were required to opt-in for the updates, after which the support team executed the remote update procedures.

4 A Shift in Secure Development Practice

Shortly after the researcher joined, one SWE from each product team was assigned to be a member of a "virtual" application security engineering team and tasked to help drive security improvements for the product. This was part of the secure software development lifecycle (S-SDLC) effort that was kicked off before we joined. The designated SWE performed security-related tasks in addition to the normal sprint work. SecSWE was assigned this role for his team.

4.1 Little Impact at First

During the first three months of the fieldwork, the only security-related work fell into two new categories of tickets created as part of the S-SDLC efforts.

- *CSF* tickets: security-related tasks guided by the NIST Cybersecurity Framework (CSF) [38].
- *ASVS* tickets: compliance with OWASP Application Security Verification Standard (ASVS) framework [40], for web facing application components.

These tickets were not included in the sprint plan. SecSWE and another developer (SWE1) were tasked to work on these tickets alongside the sprint work. Both SecSWE and SWE1 worked on these tickets individually, and the only updates about this work was provided briefly during morning scrum. These tickets were referred to as "burning cycles" and often the updates on these efforts carried little information:

- *"I knocked off a couple of CSF tickets."*
- *"I talked with <management personnel> about some CSF work and what is expected."*
- *"I will be catching up on some neglected CSF work and write up some wikis."*
- *"My changes are in PR. I will next work on ASVS tickets while I wait for reviews."*
- *"I am working on a P2 ticket and also doing some ASVS audits."*

When talking to SecSWE on how the security work is going, he remarked:

> *"I don't know. It takes a lot of work for this ASVS stuff, looking at all the code, testing, researching... I feel like we are putting all of this effort and time on this but nothing is being done about it you know."*

Although significant effort was put on resolving the CSF and ASVS tickets, we did not observe any impact on the development workflow as a whole.

4.2 Making Progress

During the third month of the fieldwork, SecSWE started to work on threat models for both products. He first shared the initial threat model for P1 with the team for feedback which garnered greater visibility on the security work in the development team as a whole and initiated discussions on the communication patterns between the different microservices in the product. SecSWE also documented the security issues in order to facilitate the pending discussions for the threat modeling work.

Prior to the threat modeling work, two security tickets had also been logged: 1) The researcher discovered that the same key pair was reused for all customers when P1 was setup as a high availability (HA) pair. 2) On further investigation, SecSWE discovered another instance of key reuse problem in establishing connections to the cloud server. The threat modeling work also initiated discussions and feedback from other SWEs concerning these issues.

Another key mismanagement issue was discovered where a private key was exposed in a publicly accessible server. The initial response from other SWEs was that this server, while Internet facing, was not advertised to the public as it was mainly used to distribute software updates. Discussions

on the potential misuse cases of this issue in particular garnered positive interest in security work with the lead SWE remarking: *"I am excited about the work SecSWE is doing."*

Security Scrum Poker. With several security tickets logged, SecSWE suggested to have a meeting specifically to discuss these tickets before the next sprint. Prior to the meeting, everyone was asked to review the tickets and corresponding wiki pages for discussion. SecSWE also introduced the DREAD risk assessment scheme [21] and the security scrum poker (akin to scrum poker) in order to assess the estimated risk of the discovered vulnerabilities. The goal was for the entire team to converge on a risk score for each ticket, discuss the rationale behind the scores in case of mismatch to clarify everyone's understanding of the issue, and ultimately use the risk scores to prioritize the security tickets.

4.2.1 Putting Security into Development Context Made Security into Development Practice

Three security scrum pokers were held during the fieldwork. In the first meeting, two SWEs tended to score lower than the others. As with scrum poker, in case of mismatch the SWEs were asked to explain the rationale behind their scores. This brought forward any misunderstanding of the discussed issue and allowed the group to clarify them. After a couple of iterations, one of the SWEs kept having varying scores and tried to move on to another ticket by agreeing with the others' scores but the lead SWE remarked: *"You cannot just do that. Either you have to defend the score or tell us why you changed your mind."* The whole team agreed that the meeting was very fruitful in clarifying their understanding of the issues and/or the proposed solutions with the SWEs remarking:

- *"That was more productive than I expected."*
- *"I really liked this session and the discussions cleared things up. I am excited to see where this effort leads."*

These discussions led to contextual analysis of the discovered issues (what is the risk in the system?). They helped uncover root causes of existing issues and bring forward discussion on potential solutions, trade-offs for alternatives, and potential road-blocks in implementing them. Importantly, **these discussions were useful to SecSWE as well.**

The discussions between SecSWE and the lead SWE led to the understanding of how and why the private key ended up in the public-facing server in the first place – it turned out that previously P1 was distributed to the customers using Preboot Execution Environment (PXE) boot over the network. Although this method had not been used for several years, it was still used internally to quickly deploy test environments. As setting up internal test environments did not require the PXE boot kickstarter script to be on a public facing server, it was subsequently moved to an internal server during the fieldwork. This task required collaboration between SecSWE, lead SWEs, as well as the networking engineers to implement, test, and deploy. For the cases of reused keys, short-term solutions of limiting users to only required commands while restricting shell access altogether were proposed. A longer-term goal to set up a per-deployment key management and distribution mechanism was also discussed. During the fieldwork only the task to research the approach was created.

After the first security scrum poker, SecSWE asked others to also report any security issues they found. During the course of the fieldwork 15 security tickets were created that were not related to ASVS or CSF. The following are the categories of vulnerabilities discovered during the fieldwork.

- Mismanagement of cryptographic keys and certificates.
- Lack of access control to remote assets
- Improper handling of passwords
- Unencrypted application update channel
- Remote code execution
- Cross-site scripting (XSS)
- Privilege escalation
- SQL and command injection
- Misconfigured SAML (Security Assertion Markup Language) authentication

These issues were discussed in at least one security scrum poker meeting. SecSWE and the researcher also developed proof-of-concept (PoC) attacks for application-level vulnerabilities such as remote code execution, XSS, Privilege escalation, and SQL and command injection which helped drive further discussions. Out of the 15 security tickets identified, 8 were approved for development after going through both the security scrum poker and the prioritization stages. Six of the approved tickets were included in a sprint plan. The researcher asked for SecSWE's opinion on the increased focus on security. The response was:

> *"I am surprised by the increased focus on security as well. They were not at all interested in these stuff before... I had already reported some of these issues before, although I didn't have time to make PoCs for it. But it's good that we have some attention now."*

4.3 Challenges in Security Ticket Prioritization

Although work was done to identify security issues, getting them prioritized for development still presented challenges.

Security tickets were not considered "real." Purely addressing existing security issues or improving security in existing code/infrastructure was not considered as "real." In one sprint planning meeting after a few security tickets were discussed and included in the sprint, the lead SWE remarked: *"Okay now let's include some real tickets in here as well."* The basis for this point of view seemed to be that security improvements made to existing features or to the infrastructure were not visible to the customers.

Security tickets had higher story points. Many security tickets were voted to have high story points and hence would not leave room to include other feature-driven tickets. The reasons for higher scores include:

- *Technical challenges:* Security tickets required more research and experimentation to figure out the most suitable solution for the product.
- *Dependencies:* Fixing existing security issues required identifying all use cases of the vulnerable feature and the impacts of the changes on the product. Finding dependencies itself was time consuming as documentations may be outdated, and additional developer and QAE testing would be needed.
- *Implied changes in processes:* SWE/support/QAE may be relying on the vulnerable features and may not want to change. SWE may need to provide viable alternatives. *"Before we move on with the fix, we need to first find out if there are undocumented use cases of these things. This is not uncommon with the support team to have some automated scripts which might rely on some access or some feature and we do not want to break them."* This could lead to additional work.

Legacy systems. Older systems already deployed at customer sites may still need to be supported. In such cases, alternative solutions needed to be provided or both new and old systems needed to be supported. Some security holes may be impossible to resolve because of initial bad design. One ticket was blocked due to this very reason as around 20 customer sites were yet to be migrated to the updated system.

Meeting Customer Requirements. Customers were unwilling to allow change of existing features. During a discussion for changing the rule specification UI, which introduced command injection vulnerability, one SWE mentioned that they had already tried to remove that feature before as the product already had an updated alternative built in. But the customers were unwilling to migrate to the new feature as it meant that they had to transition all the existing rules to the new format and they were unwilling to do so. SWE said that he already knew what this customer would say:

> *"If there are security issues then that is your problem and you need to fix it without taking away my features."*

New customer requests. During the course of a sprint, new high-priority customer tickets may be received. In such cases the security tickets would be de-prioritized, as happened to two security tickets included in the sprint plan.

4.4 Security-aware SWEs

After the introduction of security scrum pokers, there was an increase in security-related discussions outside the meetings as well. These ranged from humorous comments – *"SecSWE is not going to be happy if you do that."* or *"He is the security police now!? <laughs>"* to positive reactions for including security tickets during prioritization meetings: *"SecSWE and <the researcher> are pretty good with security."*

Security considerations in other tickets In addition to the security tickets, security considerations were made in three other feature tickets.

1. *User-side error reporting for failed certification validation.* The researcher was assigned this ticket which led to a major refactoring of the code and use of an updated single library for performing uniform certificate validation throughout P1.
2. *Enabling use of new certificate for SAML authentication without requiring application restart.* A certificate reuse misconfiguration was discvoered while working on this ticket. Code was refactored to allow proper configuration changes.
3. *Sending real-time alerts to customers.* As part of the ticket access control on cloud server was tightened to disable shell access.

Potential security issues identified. Security issues were also brought up and discussed by other SWEs.

1. An SWE discussed potential XSS vulnerabilities in another team's application while working with them, and advocated for the other team to consider upgrading a programming framework to the latest stable version.
2. Input validation was added in multiple modules proactively by SWEs working on a ticket with UI changes. Often they asked (in person or over slack) if validation code was already implemented in the module or where to look for reference validation code. In cases where validation was complicated lead SWEs proposed how the validation could be done.

   > *"Do we have any input validation code that is used both by <microservice1> and <microservice2>? If so, do you remember where it is located?"*
   >
   > *"... I know that it's not a priority for management to validate input that is supposed to be entered only by support, but it doesn't cost much."*

Security considerations in design. A feature requested by a high-priority customer required the ability to access internal configuration options otherwise hidden behind the application for an unorthodox use case of product P1. The initial design for the feature had not considered security risks with the assumption that this feature would only be accessible by the administrative account, which belongs to the support team. SecSWE pointed out that such design could potentially expose command injection and privilege escalation vulnerabilities and started a discussion on the feature, which led to the finding that the original design had overestimated the access requirements to implement the desired functionality. The initial design was then shelved with a follow up design discussion scheduled to allow time to gather information for a more secure approach.

4.4.1 What was Driving the Change

On analysis of the fieldnote data, we find that the positive shift in the development team's security awareness can be attributed to the software engineers being able to identify the applicability of the security knowledge within the context of the everyday work they performed. We observe that by working along with others in the team to apply security knowledge under the concrete context of the software products, the software engineers became attentive to security risks when similar situations were encountered later. When a considerable number of discussions had taken place on a security-related topic, the group ended up with an agreed-upon set of knowledge and the associated set of practices became the "preferred practice" for dealing with this security concern. At this stage, considering this specific aspect of code security became the group's "habit." Later if some SWE in the team needed to work on a relevant part of the code but lacked this specific piece of security knowledge, they would seek guidance from others in the team, in the same manner as they would with other types of development tasks. They then learned and executed the preferred practice of the group.

Our analysis of data shows that what was driving the positive change was the learning dynamics existent in the development team. The initial lack of visible impact from management pushing for adopting S-SDLC was because the CSF and ASVS tickets were detached from the SWEs' regular work, and thus the relevant security knowledge did not have much opportunity to be directly applied in their work. It turned out that application of knowledge was the key driver for learning in an environment like a development team. Later on when SecSWE started to use security scrum poker in the threat modeling work, and involved all SWEs in the discussions, the security knowledge became concretized and contextualized. This drove a learning cycle within the team that allowed the SWEs to start obtaining relevant security knowledge and become more security aware.

We find that understanding the learning dynamics in the development team is crucial to effectively push for secure development practices. In fact, making developers more security aware is no different than cultivating their knowledge in any other aspect of software development. **Our data indicate that, to establish a security culture in a development team, it might be helpful to follow the same learning dynamics that drive how culture forms for that community.**

5 Learning in a Development Team

The analysis of our fieldnote data yielded a model that explains the establishment and evolution of preferred practices in a development team and hence the progression of its culture. In summary, the development team is a situated learning [20] environment where the process of learning drives the creation and evolution of preferred practices. When SWEs needed assistance, they acquired the necessary knowledge from the team. As they performed their task and applied the knowledge in practice, it provided the necessary platform to further drive the process of learning and started to make contributions to the group. This process iterated over many cycles, until the group reached a point of saturation where the knowledge developed within the team was sufficient to facilitate progression in the task at hand. When this process was applied in practice, it not only led to professional growth of the SWE but also served as validation for the knowledge which then became a part of the current culture of practice.

5.1 Subject Matter Experts (SMEs)

As is common in software industry these days, the products the company built were vast entities and no single SWE knew the details of *all* aspects of a product. Multiple dimensions of knowledge were required within the development team in order to build the software, and the in-depth knowledge of each dimension was scattered between different SWEs in the team. An SWE can be the subject matter expert (SME) for some dimensions while at the same time being a novice in others.

When an SWE had the most in-depth knowledge on a topic within a development team, they were often called a subject matter expert (SME) of that particular dimension of knowledge. Although everyone in the team may have a good understanding on the topic, the SME was the one who understood the underlying details of the implementation. When an SWE started to work on a task new to them, they first went (or were directed to go) to the SME on the team. This created an implicit hierarchy within the team based on the dimension of knowledge under consideration, which facilitated the flow of knowledge within the team. This hierarchy transcended job titles. For example, despite holding a junior position in the company, a new hire who had worked on a task could immediately become the SME on certain pieces of knowledge associated with the task, and any future queries related to these pieces would first be directed towards them.

We observe the existence of SMEs throughout our data. When trying to set up a test environment for a new router device, an SWE asked the group: *"I have read through the documentation but I still cannot get it to work in our test environment. Can anyone help me out?"* He was directed to one of the network engineers: *"Normally, I just go and ask <network engineer>. I do that even before going through the documentation. 99% of the time, he knows what to do and I trust him."*

When trying to get access to a development infrastructure, the researcher asked the lead SWE: *"I need to access the CA server to test this feature. How do I get access?"* Lead SWE: *"You should go ask SWE1. He just cleaned up the access list for the CSF thing."*

When the lead SWE was asked the details of an existing

script: *"Full disclosure, I have no idea how that script works. <Former employee> implemented it and no one has had to make changes till now. But <support engineer> should provide you more information. They are the ones who use it."* In this case, although the SME is no longer within the company, the workflow established through the use of the automation script still remained and the next most knowledgeable person took responsibility of it.

Our data shows that the roles of SMEs, the knowledge on each dimension, and the preferred set of practices were not static but were developed and evolved within the development environment. When there were multiple potential SMEs on a topic, the responsibility could be passed on to the others as well. In some cases this also led to more official transfers of duties within the team. For example, when dealing with customer issues, the lead SWE was pulled into multiple meetings between the customer support team and the clients. Overwhelmed by the work, the team internally discussed the possibility of having another SWE who was working on the problem module for the past months to take over some of the client discussions, with some guidance from the lead SWE. After reaching an agreement, this was then communicated to the management for future meetings.

5.2 Establishment of Preferred Practices

The development team tended to have established preferences for activities that were carried out repeatedly. We observed team preferences for coding styles, debugging techniques, code reviews, ways of dealing with the IT department, use of scripts/tools for tasks, etc. We also observe that these preferred practices were usually tried and tested approaches of doing things within the team and were communicated to other SWEs in the team as needed. For example, preferred coding styles were communicated through the code itself while any unwarranted deviations were communicated through code reviews and reverted back to the preferred way. Any changes made to improve the existing style were also communicated through code and code review. Other preferences could be communicated mainly through discussions between the SWEs whether in a one-on-one or group setting. Usually an SWE sought help from the group using language like *"Got a second for a rubber ducky?"*, *"Can I borrow some of your time <SWE>?"* SWEs were encouraged to hold these discussions in the group chat as there could be more *"eyes"* on the problem and the solutions could be reached more quickly. These discussions also allowed for the preferred practices to evolve and improve as issues or better options were identified.

These preferred practices became a part of the group knowledge and tended to stick through generations of employees. In such cases some of the in-depth knowledge might be lost with the employee leaving but the preferred practices continued.

- *"That is a script that <former employee> developed and we still use it."*
- *"That playbook was written by a <former employee>. I know what it does but I am not sure if it uses this script internally. I would have to go read through the code but it gets the job done."*

5.3 A Situated Learning Environment

Through analysis of the field notes we find that the roles of "SME" and "learner", assumed by different SWEs for different dimensions of knowledge, drove a learning cycle within the team. This interactive activity of learning was the core process through which preferred practices were established within the team.

The pattern of learning observed here is not new. The concept of learning, not through a teacher/learner dyad, but as a situated activity where a learner not only acquires knowledge from the experts ("old-timers") and their peers but does so while participating and contributing in a community of practice is referred to as situated learning [20, 26, 32]. Learning, in this view, is not simply a process of transfer or assimilation of knowledge from the expert (SME) to learners (SWEs), but rather a generative process where each "reproduction cycle" from "learner" to "old-timer" leaves a trace in the community of practice, in both its social structure and physical, linguistic and symbolic artifacts.

The development team is a dynamic situated learning environment with a wide range of knowledge to be acquired and mastered. Based on the dimensions of knowledge under consideration, SWEs simultaneously perform multiple roles of learning practitioner, aspiring expert, status subordinate, or sole responsible agent [20]. The everyday activity of software development provided situated opportunities to learn, defining the "learning curriculum" for the task that SWEs were performing. As an SWE sought to learn from the team, different SWEs enacted different roles to drive a learning cycle to reproduce the existing culture of practice.

- *"Are you guys available for a zoom to discuss the DNS cache changes for the data viz stuff?"*
- *"Alright type gurus. I'm trying to make an interface that is a Map between two sets of constants. I'm not allowed to do what I posted above. Suggestions? . . . "*

Contradictions also arise as a part of this interactive social process as learners start to contribute. Working on resolutions to these contradictions leads to a renewed practice in the community, i.e., preferred practices are established and evolved as SWEs go through the learning cycle.

In this vein, **creating a secure development culture is the process of making secure coding practices into the preferred practices of the development team.** Thus, facilitating situated learning regarding security within the development team, is key.

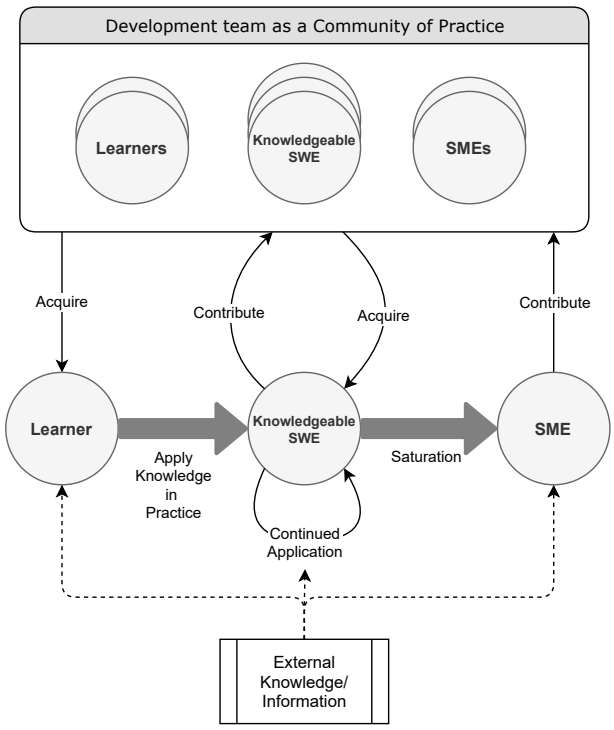


Figure 1: The Learning Cycle

5.4 The Learning Cycle

Figure 1 shows the interactions of an SWE with the development team as he/she progressed from the role of a learner to an SME. At any given time, an SWE could assume different roles for different dimensions of knowledge. For example, the SecSWE could be an SME on certain secure coding practices, but at the same time a learner on some technical details about a particular aspect of the product. External resources were accessible at anytime throughout the process of learning. We first describe the different roles an SWE could assume in this learning cycle.

- *Learner*: An SWE started out as a learner acquiring knowledge, the preferred set of practices, from the team. Learners also looked to external resources on their own, especially for a completely new aspect on which there was no existing knowledge in the team yet. Such acquisition of knowledge only made a difference in the team's practice when the SWE *applied the knowledge* in the practice, whereby he/she started to contribute to the team's knowledge and progressed in the path of professional growth.
- *Knowledgeable SWE*: The application of acquired knowledge in the context of daily practice by the SWE served an important purpose – it provided the basis to have contextual discussions whereby the SWE was able to make contributions to the group. The resulting iterations of the interactive learning process led to a convergence in understanding of the knowledge within the group, and thereby the establishment of preferred practices.
- *SME*: When the learning cycle reached saturation and no new contributions were made to the community of practice, the SWE was able to assume the role of SME. In terms of legitimate peripheral participation [20], the SWE had reached full participation for that particular dimension of knowledge space. The learning cycle then continued for that dimension with other SWEs filling in the role of learner and knowledgeable SWE.

We find that an effective learning cycle went through the following stages to create, maintain, and grow the preferred practices through multiple generations of employees.

Acquisition The most accessible and credible source for a learner was the SME on the topic of interest. They provided access to the current culture of practice to the new learner. The level of knowledge available in the team varied depending on factors such as education, prior experience, applicability in the daily work, and so on. When the knowledge within the team was sufficient, the learning cycle simply reinforced the current preferred practices, as new learners continued buying into it. In case of insufficient expertise within the team, a new knowledge requirement was created which led to individual/group research on the topic through external resources. This could also be facilitated by a new member joining the team who possessed the lacking knowledge.
Security Implication: the expertise levels of the SMEs on security *within the team* determine the team's preferred practice in secure coding.

Application Acquired knowledge needed to be applied in daily practice to drive the learning process. This was a critical step in the learning cycle; without application in daily work, the knowledge was limited to the individual SWE and never became a part of the preferred practice. On the other hand, applicability led to both individual and team growth as the applied knowledge was immediately shared to the peers through development activities like scrum meetings, design/implementation discussions, code review, testing, documentation, and so on. This provided two important driving forces that helped propagate the learning cycle: 1) a shared motivation to solve problems, and 2) the shared context of the work practices which everyone was aware of. These facilitated bi-directional discussions as opposed to a teacher-student scenario as is often perceived as how transfer of knowledge happens.
Security Implication: SWEs' security knowledge, like all other knowledge, needs to be grown with application. This works well for security, since the best time to apply security knowledge is when the code is being written (as opposed to applying security knowledge to fix vulnerabilities later on).

Contributions As SWEs put knowledge into practice, they were able to contribute to the group based on their experience and findings from daily practice. This knowledge exchange through application led to the growth and evolution of the whole team with the increase in existing knowledge on the topic. When there was an established/preferred/agreed upon knowledge base on that given topic, the knowledge in the group reached a level of saturation, and it became a part of the preferred practices.

Security Implication: When security becomes part of the preferred practice, all SWEs in the team will be security aware while writing new code. We observe that the first successful step towards implementing an effective S-SDLC and creating a security culture in the development team was the rise of security-aware SWEs. After all, if the SWEs are capable of writing secure code, it will make a real change in the final products' security. As was pointed out in prior literatures, fixing security bugs retroactively is costly and often encounters resistance from the development team [18, 25]. Companies would be better off to prevent, as much as possible, security vulnerabilities from being introduced in the first place.

6 Revisiting the Shift towards Security

We now identify the key enablers of the positive shift in secure development we observed during the fieldwork.

6.1 Setting Security as a Goal

Past experience suggested that management support was an important factor in the successful implementation of S-SDLC [19]. Our observations supported this. Due to cost in terms of time and efforts required, security was easily perceived as an "obstruction" to the daily practice of SWEs and hence the learning cycle for this dimension of knowledge did not evolve at first. We found that management played an important role to set security as a goal, making it a part of the deliverable. Doing so ensures that security knowledge is not something that overwhelms SWEs but simply applicable to daily practice, eliminating a critical barrier to drive the learning cycle.

6.2 Applying Security Knowledge in Context

Having the management directive and support for secure coding was necessary but not sufficient to eliminate the barrier to adopting secure development practices. Secure coding requires a wide range of security knowledge, and providing adequate education and awareness was pointed out as one major challenge in successfully implementing S-SDLC [19]. While the company provided SWEs virtual training for secure coding and there were also various guidelines and wiki pages the SWEs could access, applying the acquired knowledge in everyday work required expertise in both security and the contextual knowledge of the existing code base. Finding this connection was challenging for SWEs. Without application, the knowledge gained from training was at best internalized by individual SWEs, but remained detached from their daily practice. The SWEs effectively considered security-related tasks as secondary tasks, separate from their primary practice. To overcome this, a bottom up support was also needed to make real progress. In our fieldwork we found that such bottom up support happened through the learning cycle identified in the previous section. The threat modeling and associated security scrum poker meetings, which involved all SWEs, provided the opportunity for the SecSWE to put the relevant security knowledge into the concrete context of the software being built. This started the learning dynamics that enabled all SWEs to progress on the "security dimension."

6.3 The Role of Security Advocates

The work of SecSWE played an important role in facilitating the learning cycle and making security into part of the development team's preferred practices. SecSWE was a "security advocate" [15] even before the management pushed to implement S-SDLC. He worked in the development team, and was also assigned to be a part of the virtual security team, providing additional security resources. Analyzing our data, we find that this structure added more value to security advocacy, making other SWEs more receptive to his advice as they started to consider it "part of his job." Working on the same team provided an important factor in demonstrating the applicability of the security knowledge in the context of the daily practice. This facilitated SecSWE to contribute knowledge as applicable to daily practice, helping to drive a productive learning cycle, which was beneficial to both the rest of the team and SecSWE himself. Through this interactive learning process, SecSWE was able to better understand the necessary details of the product which allowed him to apply his security knowledge in a more context-aware manner. Further iterations of this learning cycle led to more security-aware SWEs in the team.

7 Limitations

Our work is limited by a few factors. First, our findings are based on the fieldwork data collected by a single researcher. Although the researcher had prior training and experience in conducting participant observation research, the collected data are shaped by the researcher's positionality (his age, gender, position in the company, and so forth). For example, the researcher did not have as many interactions with customer service and upper management because of his position in the company. However, the researcher did build an overall understanding of the company during the research, and the results were extensively discussed with the broader research group during analysis to better account for any inherent biases in the

data. Second, our findings are based on the observations of a single company with a particular size and structure. Although we believe the development team is representative of one in a mid-sized software development company, the specific challenges of adopting secure development practices and how they were/were not overcome may not be directly applicable and generalizable to every company. As such, the model of how a culture is developed within a software development team might not be comprehensive. Nevertheless, during data analysis, the team paid particular attention to how results related to common problems faced in security and software development to ensure that the findings could be relevant to other companies.

8 Recommendations for Companies

Our findings suggest a potentially useful strategy for a small to medium sized company. Having a security expert as a part of the development team, participating and advocating for security at every stage of the development process, is beneficial in starting a security culture. This not only helps cultivate security-aware developers, but also helps the security expert identify security issues and collectively converge to secure practices that are best suited for the project at hand. Development of the relevant security knowledge in conjunction with the regular software development skills promotes secure coding practices which, overtime, become a part of the team culture. Our research also observed the effect management had in facilitating the positive shift. Even though the initial efforts focused on the compliance tickets were not effective, the fact that management made security an explicit goal provided the opportunity for the security advocates to experiment different strategies that eventually led to positive results.

9 Related Work

Our fieldwork was conducted in the backdrop of the company starting to implement a secure development lifecycle, a concept first articulated by Howard and Lipner [19]. This seminal work highlighted the importance of education and training in creating S-SDLC. Our findings further indicate that understanding the learning dynamics, in particular how preferred practices are established within a software development team through the situated learning framework, can be instrumental in creating positive changes in secure development.

There is a long line of study on developers' role in software security. Some used psychological techniques [24]. Others used surveys and interviews [3, 5, 13, 22, 28, 35] as well as study of code artifacts [3, 22]. More recently, researchers have used secure coding competitions [27, 31] and controlled experiments [2–4, 11, 23] to study the problem. Our work is unique in that we use long-term participant observation conducted in a real company. The longitudinal study based on real-world observations allows us to obtain deep insights that are otherwise hard to come out through snapshots-in-time study or self-reported data.

Palombo et al. [25] used ethnographic methods to study a software company's secure development processes. The authors indicated that a co-creation model where security experts working inside the development team could produce positive changes in secure development processes. Our work revealed the role of learning dynamics in pushing for positive shift in adopting secure development processes. The role situated learning plays in starting a secure development culture is consistent with the co-creation model.

The SecSWE in our study can be viewed as a "security advocate," which has been extensively discussed in recent studies [14–16]. Our findings on the role of team culture in security awareness of SWEs echoes that from prior studies. Assal and Chiasson [5] explored how security best practices are integrated into the software development lifecycles and found that company culture is an influential factor in adoption of security practices. Haney et al. [17] carried out in-depth interviews to understand cryptographic development and testing practices in organizations and found that rigorous secure development and testing practices are guided by a strong security culture within organizations. They also identify that security experts within the team are critical influences in the security culture of an organization and in supporting less-experienced personnel. Our findings confirm the important role security advocates play in starting a security culture, and further provide guidance on how to make security advocates' work effective, through understanding the underlying learning dynamics that drive the formation of a development team's culture.

There are also past work that examined the effect of learning from experience in software development [9], and work that analyzed open-source software development using the situated learning framework [8]. Our work focuses on secure development, and our research findings are consistent with these earlier works which focused on learning's role in software development in general.

10 Conclusion

We present an ethnographic study of secure development processes in a software company. Our research was able to observe the unfolding of implementing a secure development life cycle in the company. Data analysis shows that a positive shift in developers' security awareness resulted from underlying situated learning dynamics, where security knowledge is constantly applied in the concrete work of the development team. This process drives the establishment of secure coding practices as the preferred practices of the team, essentially establishing a secure development culture. We find that a security expert working within the development team could be instrumental in driving this positive shift.

Acknowledgments

We thank Raj Rajagopalan, John McHugh, and the anonymous reviewers for numerous valuable comments on an earlier version of this paper. We owe gratitude to the company, and its employees who participated in the study. This research is supported by the U.S. National Science Foundation under Grant No. 1801633. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Y. Acar, S. Fahl, and M. L. Mazurek. You are not your developer, either: A research agenda for usable security and privacy research beyond end users. In *2016 IEEE Cybersecurity Development (SecDev)*, pages 3–8, 2016.

[2] Yasemin Acar, Michael Backes, Sascha Fahl, Simson Garfinkel, Doowon Kim, Michelle L Mazurek, and Christian Stransky. Comparing the usability of cryptographic APIs. In *IEEE Symposium on Security and Privacy*, San Jose, CA, USA, 2017.

[3] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L Mazurek, and Christian Stransky. You get where you're looking for: The impact of information sources on code security. In *IEEE Symposium on Security and Privacy*, San Jose, CA, USA, 2016.

[4] Yasemin Acar, Christian Stransky, Dominik Wermke, Michelle L Mazurek, and Sascha Fahl. Security developer studies with github users: Exploring a convenience sample. In *Thirteenth Symposium on Usable Privacy and Security ({SOUPS} 2017)*, pages 81–95, 2017.

[5] Hala Assal and Sonia Chiasson. Security in the software development lifecycle. In *14th Symposium on Usable Privacy and Security*, Baltimore, MD, USA, 2018.

[6] H Russell Bernard, Amber Wutich, and Gery W Ryan. *Analyzing qualitative data: Systematic approaches*. SAGE publications, 2016.

[7] Kathleen M. DeWalt and Billie R. DeWalt. *Participant Observation: A Guide for Fieldworkers*. Lanham: AltaMira Press, second edition, 2011.

[8] Kasper Edwards. Epistemic communities, situated learning and open source software development, 2001.

[9] Wai Fong Boh, Sandra A Slaughter, and J Alberto Espinosa. Learning from experience in software development: A multilevel analysis. *Management science*, 53(8):1315–1331, 2007.

[10] David Geer. Are companies actually using secure development life cycles? *Computer*, 43(6):12–16, 2010.

[11] Peter Leo Gorski, Luigi Lo Iacono, Dominik Wermke, Christian Stransky, Sebastian Möller, Yasemin Acar, and Sascha Fahl. Developers deserve security warnings, too: On the effect of integrated security advice on cryptographic API misuse. In *14th Symposium on Usable Privacy and Security*, Baltimore, MD, USA, 2018.

[12] M. Green and M. Smith. Developers are not the enemy!: The need for usable security apis. *IEEE Security Privacy*, 14(5):40–46, 2016.

[13] Matthew Green and Matthew Smith. Developers are not the enemy!: The need for usable security APIs. *IEEE Security & Privacy*, 14(5):40–46, 2016.

[14] Julie M. Haney and Wayne Lutters. Security awareness in action: A case study. In *Workshop on Security Information Workers, USENIX Symposium on Usable Privacy and Security*, Santa Clara, CA, USA, 2019.

[15] Julie M. Haney and Wayne G. Lutters. "It's scary... it's confusing... it's dull": How cybersecurity advocates overcome negative perceptions of security. In *14th Symposium on Usable Privacy and Security*, Baltimore, MD, USA, 2018.

[16] Julie M. Haney and Wayne G. Lutters. Motivating cybersecurity advocates: Implications for recruitment and retention. In *Computers and People Research Conference*, Nashville, TN, USA, 2019. Association for Computing Machinery.

[17] Julie M. Haney, Mary Theofanos, Yasemin Acar, and Sandra Spickard Prettyman. "We make it a big deal in the company": Security mindsets in organizations that develop cryptographic products. In *14th Symposium on Usable Privacy and Security*, Baltimore, MD, USA, 2018.

[18] Bill Haskins, Jonette Stecklein, Brandon Dick, Gregory Moroney, Randy Lovell, and James Dabney. 8.4.2 error cost escalation through the project life cycle. *INCOSE International Symposium*, 14:1723–1737, 06 2004.

[19] Michael Howard and Steve Lipner. *The security development Lifecycle*, volume 8. Microsoft Press Redmond, 2006.

[20] Jean Lave, Etienne Wenger, et al. *Situated learning: Legitimate peripheral participation*. Cambridge university press, 1991.

[21] JD Meier. *Improving web application security: threats and countermeasures*. Microsoft press, 2003.

[22] Sarah Nadi, Stefan Krüger, Mira Mezini, and Eric Bodden. Jumping through hoops: Why do java developers struggle with cryptography apis? In *Proceedings of the 38th International Conference on Software Engineering*, ICSE '16, page 935–946, New York, NY, USA, 2016. Association for Computing Machinery.

[23] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, Marco Herzog, Sergej Dechand, and Matthew Smith. Why do developers get password storage wrong? a qualitative usability study. In *ACM SIGSAC Conference on Computer and Communications Security*, Dallas, Tex, USA, 2017.

[24] Daniela Oliveira, Marissa Rosenthal, Nicole Morin, Kuo-Chuan Yeh, Justin Cappos, and Yanyan Zhuang. It's the psychology stupid: how heuristics explain software vulnerabilities and how priming can illuminate developer's blind spots. In *30th Annual Computer Security Applications Conference*, New Orleans, LA, USA, 2014.

[25] Hernan Palombo, Armin Ziaie Tabari, Daniel Lende, Jay Ligatti, and Xinming Ou. An ethnographic understanding of software (in)security and a co-creation model to improve secure software development. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*, pages 205–220. USENIX Association, August 2020.

[26] Barbara Rogoff. Developing understanding of the idea of communities of learners. *Mind, culture, and activity*, 1(4):209–229, 1994.

[27] Andrew Ruef, Michael Hicks, James Parker, Dave Levin, Michelle L Mazurek, and Piotr Mardziel. Build it, Break it, Fix it: Contesting secure development. In *2016 ACM SIGSAC Conference on Computer and Communications Security*, Vienna, Austria, 2016.

[28] Adam Shostack, Matthew Smith, Sam Weber, and Mary Ellen Zurko. Empirical Evaluation of Secure Development Processes (Dagstuhl Seminar 19231). *Dagstuhl Reports*, 9(6):1–25, 2019.

[29] James P. Spradley. *Participant Observation*. Holt, Rinehart, and Winston, 1980.

[30] David R Thomas. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation*, 27(2):237–246, 2006.

[31] Daniel Votipka, Kelsey Fulton, James Parker, Matthew Hou, Michelle L. Mazurek, and Michael Hicks. Understanding security mistakes developers make: Qualitative analysis from Build It, Break It, Fix It. In *29th USENIX Security Symposium*, Boston, MA, USA, 2020.

[32] Etienne Wenger. *Communities of practice: Learning, meaning, and identity*. Cambridge university press, 1999.

[33] Glenn Wurster and P. C. van Oorschot. The developer is the enemy. In *Proceedings of the 2008 New Security Paradigms Workshop*, NSPW '08, page 89–97, New York, NY, USA, 2008. Association for Computing Machinery.

[34] Shundan Xiao, Jim Witschey, and Emerson Murphy-Hill. Social influences on secure development tool adoption: Why security tools spread. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, page 1095–1106, New York, NY, USA, 2014. Association for Computing Machinery.

[35] J. Xie, H. R. Lipford, and B. Chu. Why do programmers make security errors? In *2011 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 161–164, 2011.

[36] Build security in maturity model | BISIMM. https://www.bsimm.com/.

[37] Code quality and security | SonarQube. https://www.sonarqube.org/.

[38] Framework for improving critical infrastructure cybersecurity. https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.04162018.pdf.

[39] The legion of the bouncy castle. https://www.bouncycastle.org/.

[40] OWASP ASVS. https://owasp.org/www-project-application-security-verification-standard/.

[41] Planning poker - wikipedia. https://en.wikipedia.org/wiki/Planning_poker.

[42] The security mindset. https://www.schneier.com/blog/archives/2008/03/the_security_mi_1.html.

[43] Software assurance maturity model (SAMM). https://www.opensamm.org/.

[44] Software composition analysis | Black Duck Software. https://www.blackducksoftware.com/.

A Appendix: Codebook

Coding of the fieldnote data followed the general guidelines of inductive approach [30] and grounded theory [6]. It was an iterative process and started after the first three months of the field work. The research team held weekly meetings to discuss and reflect on the data collected. Themes and patterns emerged from those discussions and various codes were used to tag the content in the raw fieldnote. Coding was done by the embedded researcher only, to protect the privacy of participants. Below is the list of codes used.

- Bug discovery
- Bug discovery:internal
- Communication issue
- Compliance:asvs
- Compliance:csf
- Compliance:csf:thirdparty
- Compliance:encryption
- Compliance:phishing
- Cross product issue
- Customer pressure
- Feature pressure
- Forgotten issue
- Ignored issue
- Infra
- Infra:legacy
- Infra:security
- Learn
- Learn:best practice
- Learn:figure out
- Learn:peer programming
- Learn:review
- Policy change
- Preferred practice:code
- Preferred practice:support
- Preferred practice:workflow
- Remote work issues
- SME
- SME:handover
- SME:new
- Secure development
- Security-aware
- Threat modeling
- Threat modeling:dread
- Training
- Workflow change

Examining the Examiners: Students' Privacy and Security Perceptions of Online Proctoring Services

David G. Balash[‡], Dongkun Kim[‡], Darika Shaibekova[‡]
Rahel A. Fainchtein[§], Micah Sherr[§], and Adam J. Aviv[‡]
[‡] *The George Washington University,* [§] *Georgetown University*

Abstract

In response to the Covid-19 pandemic, educational institutions quickly transitioned to remote learning. The problem of how to perform student assessment in an online environment has become increasingly relevant, leading many institutions and educators to turn to online proctoring services to administer remote exams. These services employ various student monitoring methods to curb cheating, including restricted ("lockdown") browser modes, video/screen monitoring, local network traffic analysis, and eye tracking. In this paper, we explore the security and privacy *perceptions* of the student test-takers being proctored. We analyze user reviews of proctoring services' browser extensions and subsequently perform an online survey ($n = 102$). Our findings indicate that participants are concerned about both the amount and the personal nature of the information shared with the exam proctoring companies. However, many participants also recognize a trade-off between pandemic safety concerns and the arguably invasive means by which proctoring services ensure exam integrity. Our findings also suggest that institutional power dynamics and students' trust in their institutions may dissuade students' opposition to remote proctoring.

1 Introduction

In the past decade colleges and universities have steadily expanded online course offerings [19]. The Covid-19 pandemic has significantly accelerated that pace, as in-person classes were quickly replaced with virtual instruction [2]. With the increase in online education, academic integrity issues surrounding how students complete online exams led many educators to utilize *remote proctoring services* [7].[1] A 2020 EDUCAUSE poll found that more than half of higher education institutions use remote proctoring services and another 23% are either planning for or considering their use [9].

Remote proctoring services are offered by a number of companies, including popular vendors such as Respondus [29], Proctorio [25], and ProctorU [26]. Many remote proctoring services require students to install a browser extension that "locks down" their browser, preventing navigation to other sites during exam time. However, more invasive monitoring may also include webcams, screen sharing, the use of a live (human) proctor, and even automated monitoring techniques such as eye tracking and network traffic analysis.

There is evidence of higher rates of academic integrity violations for online exams [18, 22], and some argue that online proctoring is an effective tool to curb cheating [14]. However, this can come at the expense of increased test anxiety and diminished student performance [6]. Importantly, the privacy policies and practices of these services, and of online education, generally, significantly impact students and their privacy rights [4]. While concerns over the privacy and the ethics of online exam proctoring have led several institutions (cf. [1, 17]) to discontinue their contracts with online proctoring services, there is little research on the privacy perceptions and understandings *of the student test-takers* who undergo remote proctoring. In this paper we endeavor to answer the following research questions about privacy in the setting of online proctoring services:

RQ1 What are students' perceptions and understandings of online proctoring services?
RQ2 What are students' privacy concerns regarding the use of online proctoring software?
RQ3 What are students' security concerns regarding the use of online proctoring software?

[1] Remote proctoring services are sometimes called online proctoring services, or more simply, online proctoring. We use these terms interchangeably in this paper.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

We first reviewed eight online proctoring services' Chrome browser extensions. Based on the number of user reviews, we observed explosive growth of online proctoring since the start of the Covid-19 pandemic (720 %). Qualitative analysis of the user reviews revealed a number of privacy concerns, including providing personal identifiable information to verify students' identities, live-proctors viewing webcams, local network monitoring, and screen sharing.

We subsequently developed an online survey to further explore privacy issues with $n = 102$ student participants who took an online proctored exam. Only 39 % of participants *agreed* or *strongly agreed* that they prefer an online proctored exam, and participants expressed many of the same privacy concerns as found in user reviews, particularly around the process of identity verification. They also expressed concern for installing proctoring software. A little more than half were at least *somewhat*, *moderately* or *extremely* concerned about installing proctoring software on their personal computers, and 52 % *agreed* or *strongly agreed* that exam proctoring was too privacy invasive. Participants were more comfortable with lockdown browsers, keyboard restrictions and even a live proctor while being monitored but expressed discomfort with screen, webcam or microphone recording. They were least comfortable with browser history monitoring.

Despite concerns, many participants noted a privacy-benefit trade-off in their qualitative responses, recognizing that taking exams online was more convenient and safe during the Covid-19 pandemic. At the same time, many participants also indicated that they did not believe online proctoring prevents academic dishonesty: 61 % noted that they *agreed* or *strongly agreed* that they could still cheat (if they wanted to).

We also found that power dynamics shaped students' perceptions of online proctoring. Students reported that for 97 % of remotely proctored exams, the proctoring was required by their instructor or institution. The obligatory monitoring and its backing by academic institutions may explain why many participants are able to contextualize their privacy exposure. Some participants noted that their trust in the proctoring services was due in part to their belief that their institution would not harm their security or privacy.

Given our findings, we present a number of recommendations for educators. These include acknowledging students' concerns regarding remote proctoring services, better communicating the privacy and security implications of using these services, presenting a clear rationale for using the selected proctoring system, providing some form of consent and notice to students before online proctored exams, and providing clear instructions and/or assistance in removing invasive monitoring software following an exam.

2 Background and Related Work

Online exam proctoring services enable students to complete an exam (or other coursework) online while being proctored remotely. When taking an online exam, students may be required to install software to assist in confirming their identity, monitoring their behavior, and preventing their access to unauthorized resources. Monitoring may include the use of the webcam and microphone, sharing computer screens, monitoring the network, eye tracking, or other behavioral tracking. Some services use a live (human) proctor to observe the student. While there are other mechanisms for remote examination, such as taking an exam using video conferencing (e.g., Zoom), this study is focused on remote online proctoring services that provide a more comprehensive observation using browser plugins and/or standalone software, as well as student identify verification. Herein, when we refer to "online exam proctoring" or an "online proctored exam" we are specifically referring to services as described above.

Despite considerable media attention [11, 13, 23, 31], student *perceptions* of online proctoring services have been understudied. We identified one recent study by Kharbat and Abu Daabes which finds high levels of privacy concerns in the UAE when using online proctoring systems [16]; we find similar results. However, unlike Kharbat and Abu Daabes, we focus on participants' security and privacy concerns, and how they compare with the risks we identified through our own analysis of these tools. A recent manuscript by Cohney et al. explores privacy risks of online proctoring services [4] but instead focuses on perceptions of university administrators and faculty; we focus on the student perspective.

Cheating during online exams has been investigated, with sometimes contradictory findings. Watson and Sottile found that students indicated they would be 4x more likely to cheat in online classes, but more readily during in-person exams [32]. Lanier compared rates of academic dishonesty at a university that offered both online and in-person learning, finding more cheating in the online courses [18]. However, Grijalva et al. found that the rate of cheating in online classes resembles that of traditionally proctored exams [10].

Hylton et al. examined whether webcam-based monitoring had a deterrent effect [14]. They found no statistically significant difference in exam scores between students who were and were not monitored, but report that non-proctored students took longer to complete exams and perceived they had more opportunity to cheat. Similarly, Rios and Liu also found little difference in exam performance on low-stakes exams, suggesting that rates of cheating are also similar between low-stakes proctored and non-proctored exams [30].

In contrast, Daffin and Jones found that student performance was generally 10-20% higher on online psychology exams that did not use online proctoring services [6]. Goedl and Malla also found that student performance was significantly greater without online proctoring. However, like Hylton et al., they found students consistently took less time to complete their exams when proctored [8]. In these studies, it is unclear whether the differences in completion time and performance were due to (1) the online proctoring acting as

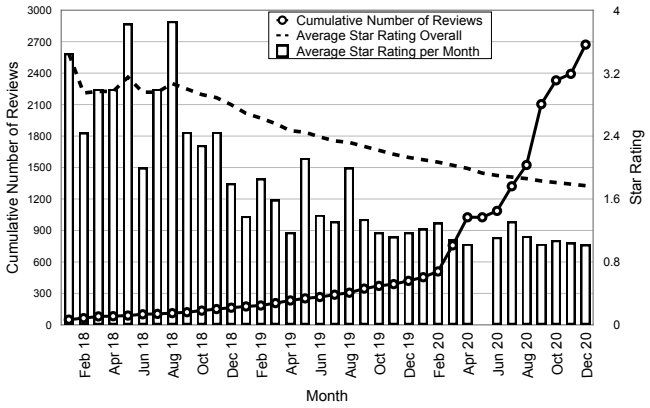


Figure 1: Number of Chrome Web Store reviews and star ratings for exam proctoring browser extensions ($n = 8$).

a deterrent to curb cheating that would otherwise result in higher test scores or (2) a psychological effect of the presence of the remote monitoring. Woldeab and Brothen addressed this more specifically and found that for students with trait test anxiety, exam-time stress was more closely correlated with poorer performance in online proctored exams [33].

In two recent opinion articles, Coghlan et al. highlight ethical considerations when integrating machine learning and artificial intelligence techniques into online proctoring services [3], and Swauger opines that the algorithms that underpin online monitoring have been shown to "[reinforce] white supremacy, sexism, ableism, and transphobia" and that proctoring services inherit these traits [31]. These types of concerns, particularly those of online proctors' inadequate accessibility and protection of student privacy, have led to the cancellation or discontinuation of contracts with online proctoring vendors at the University of Illinois [17] and the University of California, Berkeley [1].

3 Browser Extension and Privacy Policies

As an initial investigation into online proctoring services, we conducted a study of Chrome Web Store reviews of online proctoring services' browser extensions. These extensions are often required to be installed as a prerequisite to taking online proctored exams. We analyzed user reviews from the Chrome Web Store posted between October 2015 and December 2020 for eight browser extensions. We also analyzed the privacy policies of 25 proctoring services, as reported on their websites with respect to the kinds of information collection and monitoring practices. This analysis informs the development of the online survey discussed in Section 4.

Growth in Online Proctoring We first analyzed the number of reviews over time, dating back to January 2018. (See Figure 1.) While there is a steady rise in the number of reviews, starting in January 2020 (the beginning of the Covid-19 pandemic) the growth in reviews greatly increased. By the end of 2020, exam proctoring browser extensions experienced an 8.2x (720 %) increase in the number of reviews, totaling 2,348 reviews. In the prior two years (2018, 2019), only 292 reviews appeared on the web store, strongly suggesting that the Covid-19 pandemic has led to a large expansion of students who are taking remotely proctored exams. This confirms a recent poll by Grajek that found that more than half of colleges and universities make use of online proctoring [9].

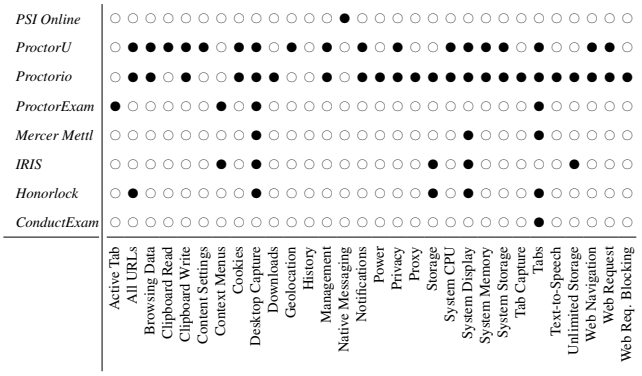
Interestingly, as shown in Figure 1, there is a noticeable decline in the average star rating that coincides with the start of the pandemic (and the growth in popularity of online proctoring services). Remarkably, by the end of 2020, the average rating fell to just 1.02 (the lowest possible rating is 1).

Analysis of Reviews We analyzed a total of 613 reviews that were written between August 2015 and October 2020 for the browser extensions offered by ConductExam [5], Honor-

Table 1: The number of results found for each URL match pattern in the Honorlock browser extension manifest file. To obtain this data we used the Google site operator (e.g., `site:http://*/courses/*/quizzes/*`) in February 2021.

Pattern	*Matching URLs*
`http://*/courses/*/quizzes/*`	8
`https://*/courses/*/quizzes/*`	99,300
`http://*/courses/*/quizzes/*/take?user_id=*`	8
`https://*/courses/*/quizzes/*/take?user_id=*`	8
`https://*/courses/*/quizzes*`	9
`*://*/d2l/lms/quizzing/*`	8
`*://*/webapps/assessment/*`	316,000
`*://*/ultra/courses/*`	9
`http://*/courses/*/quizzes`	183,000
`https://*/courses/*/quizzes`	240,000
`http://*/courses/*/quizzes/*/take`	9
`http://*/courses/*/quizzes/*/take/questions/*`	8
`https://*/courses/*/quizzes/*/take`	231,000
`https://*/courses/*/quizzes/*/take/questions/*`	8
`*://*/webapps/assessment/*`	316,000
`*://*/d2l/lms/quizzing/*`	8
Total Matches	1,385,383

Table 2: Permission access of browser extensions.



lock [12], IRIS [15], Mercer Mettl [21], ProctorExam [24], Proctorio [25], ProctorU [26] and PSI Online [27]. A primary coder crafted a codebook by coding a random sample of (up to) 100 reviews per extension. (Some extensions had fewer than 100 reviews.) Using the codebook, a secondary coder coded all reviews over several rounds, providing feedback on the codebook and iterating with the primary coder until inter-coder agreement was reached (Cohen's $\kappa > 0.7$).

We find that 83 % ($n = 510$) of users shared negative reviews. For instance, a user stated, "Just an absolute nightmare to use," and, "It is a small wonder how they convinced all these companies to use it for their online exams." The most prevalent concern was the 55 % ($n = 335$) who mentioned concerns about their privacy. For example, "I'm not letting some random person have control over facial recognition of me and scan the inside of my home." A number of reviews noted positive experiences ($n = 73$; 12 %), e.g., "It has gotten the job done, and I have never had any problems with it." A few reviews ($n = 60$; 10 %) mentioned that the use of proctoring services was required by their institution.

Monitoring Techniques and Scope We also extracted *manifest* files from each extension, which describe the permissions (or access level) for the extension and on which web pages the extension is active. Pages on which the extension is active are indicated by a list of URL match patterns, with * indicating a wild card. We found that the extensions' URL matching can be quite broad. For example, in Table 1, we report the number of Google search results that match each URL pattern specified by Honorlock's browser extension. These URL patterns match a wide variety of URLs, most likely associated with online course content hosted through Blackboard[2] or Canvas.[3] However, generic URL patterns can match other URLs (e.g., any URL that has `/courses/` followed by `/quizzes/`), activating the browser extension regardless of whether the student is taking an exam.

This can be problematic for student privacy, beyond the duration of the exam, as these browser extensions request many browser permissions in order to conduct monitoring. Table 2 reports the permission requests for the eight proctoring browser extensions. All but two extensions request multiple permissions. ProctorU and Proctorio request the most, with Proctorio requesting 22 different permissions, which could be active when visiting any page matching a URL pattern.

Privacy Policies In addition to viewing permissions in the manifest file, we also reviewed the privacy policies of 25 exam proctoring services. See Question **Q6** for a full list; not all had browser extensions in the web store. Figure 2 presents the number of exam proctoring services (x-axis) that disclose certain data collection practices (y-axis). All 25 discuss setting cookies, collecting IP addresses, and accessing the webcam, and all but one note access to a photo ID to verify identity. Many policies also mention that the software will request access to the microphone, screen recordings, or collect other kinds of biometric information. Notably, 18 state that they share information with third parties.

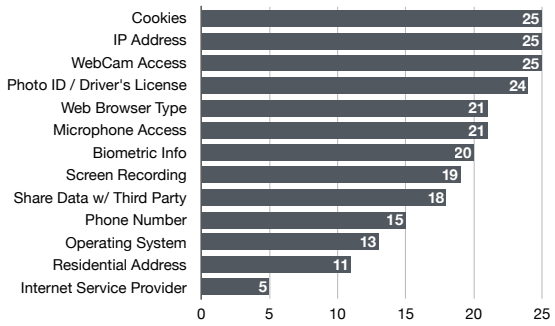


Figure 2: Data collection disclosed by exam proctoring services in their privacy policies ($n = 25$).

4 Survey Methodology

We conducted an online survey to evaluate the security and privacy concerns of student test-takers who are remotely proctored. The design of our study is informed by our preliminary analysis of the browser extensions and the privacy policies (see Section 3), and in what follows, we describe the survey's procedures, recruitment, limitations, and ethics. Survey results are presented in Section 5.

4.1 Study Procedure

To ensure that participants had taken at least one online proctored exam, we used a two-part structure with an initial *screening survey* in which qualified participants were then asked to participate in the *main study*. The full text of the screening survey and main study can be respectively found in Appendices A.1 and A.2.

Screening Survey We used the following two inclusion criteria to screen participants for the main study: (1) the participant is familiar with online exam proctoring and (2) the participant has taken an online proctored exam.

In the screening survey we also asked participants to describe their overall experience taking online proctored exams and to provide demographic information such as age, identified gender, education, and technical background. Participants also answered the Internet Users' Information Privacy Concerns (IUIPC) questionnaire [20] to provide insights into their privacy concerns.

[2] https://www.blackboard.com
[3] https://www.instructure.com/canvas

Main Study The main study consisted of the following:

1. Informed Consent: Participants were asked to consent to the study. The consent included that participants would answer questions about their awareness and concerns about online exam proctoring services.
2. Awareness and Exposure: Participants were asked to report their experiences with online exam proctoring, including the number of exams taken, the nature of the exams, the proctoring service(s) used, and if they were required to take the exam. Participants were also asked if the online proctoring service provided any necessary accommodations or other modifications based on their needs as a test taker, and if they experienced any technical difficulties during the exam. These questions were informed by the browser extension reviews. Questions: **Q1**-**Q17**.
3. Proctoring Methods: Next, participants were asked about their level of comfort with specific monitoring methods used by proctoring services, such as eye movement tracking, video monitoring, and internet activity monitoring, and if these monitoring methods were necessary. The list of these methods were informed by the analysis of the browser extensions and privacy policies. Questions: **Q17**-**Q28**.
4. Proctoring Effectiveness: To determine the perceived effectiveness of online exam proctoring we asked if participants were less likely to cheat and if they believed it is still possible to cheat on an exam even with the monitoring methods employed by online exam proctoring services. Additionally, participants were asked if they had been accused of cheating by exam proctoring software and, if so, which specific methods such as eye movement tracking, screen recording, or internet activity monitoring was used to detect cheating. Participants could choose to not answer these questions. Questions: **Q29**-**Q33**.
5. Privacy Concerns: Participants were asked to evaluate their concern regarding sharing information with online exam proctoring companies, whether the proctoring service was a reasonable trade-off between personal privacy and the integrity of the exam, and whether online exam proctoring was a good solution for monitoring remote examinations. Questions: **Q34**-**Q39**.
6. Proctoring Software: Finally, participants were asked about the installation of exam proctoring software, what the software did, and their level of concern about the software. Questions: **Q40**-**Q50**.

4.2 Recruitment and Demographics

We initially recruited 27 participants by posting an advertisement on Reddit via the subreddit *SampleSize*[4] between November 14, 2020 and December 2, 2020. Note that participants recruited via Reddit did not take the screening survey, but rather the pre-survey questions were included in the main study. We excluded responses that did not meet the screening criteria.

Table 3: Demographic and IUIPC data collected at the end of the screening survey.

		Screening (n = 178)		**Main Study** (n = 102)	
		n	**%**	**n**	**%**
Gender	Woman	85	48	47	46
	Man	85	48	52	51
	Non-binary	7	4	2	2
	No answer	1	1	1	1
Age	18–24	124	70	73	72
	25–34	39	22	22	22
	35–44	9	5	5	5
	45–54	4	2	2	2
	55+	2	1	0	0
		Avg.	**SD**	**Avg.**	**SD**
IUIPC	Control	5.9	0.8	6.0	0.8
	Awareness	6.5	0.6	6.5	0.6
	Collection	5.7	1.0	5.7	1.1
	IUIPC Combined	6.0	0.6	6.0	0.7

We were not able to find a sufficiently large sample on Reddit, and so we recruited additional participants on *Prolific*[5] between December 18, 2020 and December 28, 2020. As a part of the screening survey we recruited 150 participants. Using their ProlificIDs, we re-recruited 75 of the participants who met the criteria for participation in the main study.

Participants who completed the Reddit survey were given the opportunity to enter a drawing for a $50 USD Amazon gift card with a 1 in 27 chance of winning. On average, it took 27.2 minutes (SD=24.5) to complete the Reddit survey. Participants who completed the screening survey received $0.50 USD. On average, it took 4.2 minutes (SD=2.7) to complete the screening survey and 15 minutes (SD=7.1) to complete the main study. Participants who completed the main study received $3.50 USD.

Seventy-two percent of main study participants were between 18–24 years old, 22 % were between 25–34 years old, and 7 % were 35 years or older. The identified gender distribution for the main study was 51 % men, 46 % women, and 3 % non-binary or did not disclose gender. Participant characteristics are presented in Table 3, and additional demographic information can be found in Appendix B. In total, $n = 102$ participants were recruited for the main study.

[4] https://www.reddit.com/r/SampleSize

[5] https://www.prolific.co

4.3 Ethical Considerations and Limitations

The study protocol was approved by our Institutional Review Board (IRB) with approval number NCR202908, and all collected data is associated with random identifiers. Throughout this process we considered that many participants may not want to share their perceptions of whether proctoring services are effective at preventing academic dishonesty, and so we made those questions optional.

Our study is limited in its recruitment, particularly to Prolific and Reddit users residing in the U.S. We cannot claim full generalizability of the results. Despite this limitation, prior work [28] suggests that online studies about privacy and security behavior can approximate behaviors of populations.

We are also limited by the fact that this study relies on self-reported behavior. We cannot verify that the participants actually experienced an online proctored exam, which is why we used a screening survey. Finally, responses can suffer from social desirability and response bias, leading participants to over describe their awareness of online exam proctoring as they may believe that this is the expectation of the researchers. Such biases may be most present when participants indicate concerns and indicate they are less likely to cheat on an exam.

5 Results

We organize our results according to our research questions. We first present our findings concerning participants' perceptions and understanding of online exam proctoring (**RQ1**), and then describe participants' privacy concerns regarding online exam proctoring (**RQ2**). Finally, we discuss participants' understanding of exam proctoring software and their concerns about such software (**RQ3**).

For all qualitative findings, we used a pair of primary coders from the research team, each of whom crafted a codebook and identified descriptive themes by coding each question. A secondary coder coded a 20 % sub-sample from each of the free-response questions over several rounds, providing feedback on the codebook and iterating with the primary coder until inter-coder agreement was reached (Cohen's $\kappa > 0.7$).

5.1 RQ1: Perceptions and Understanding

As part of **RQ1**, we seek to measure (1) student perceptions of online proctoring and (2) their understanding of the methods used by online proctoring services to monitor exams.

Experience with Exam Proctoring Nearly half ($n = 49$; 48 %) of respondents had taken five or more online proctored exams, 38 % ($n = 39$) had taken between two to four (inclusive), and a mere 14 % ($n = 14$) of participants had only taken a single online-proctored test (**Q1**). The online proctoring service *Respondus* was the most used ($n = 20$; 20 %), followed by *Proctorio* ($n = 13$; 13 %), and *ProctorU* ($n = 10$; 10 %)

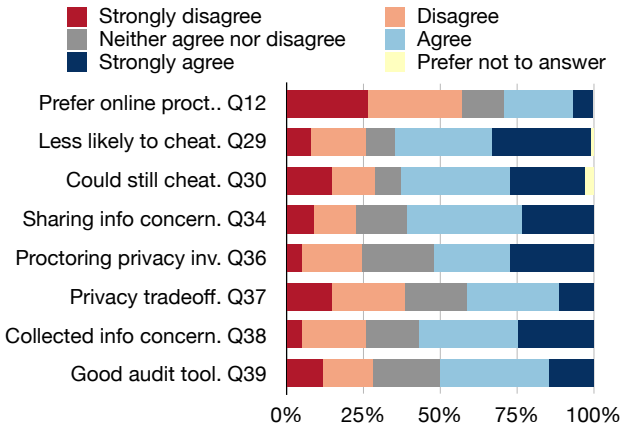


Figure 3: Impressions of online proctoring services.

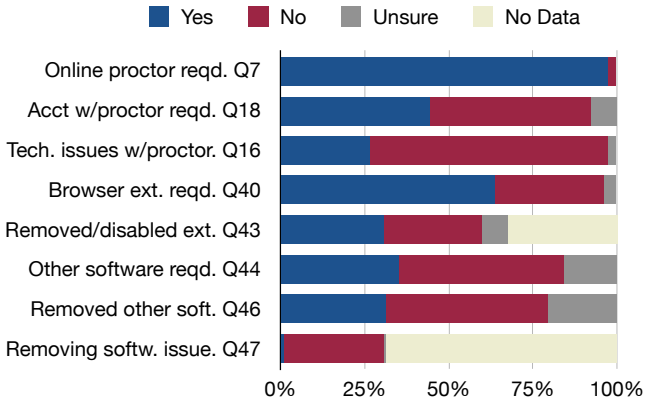


Figure 4: Encountered exam requirements.

(**Q6**) (see Figure 10 in Appendix B). This generally conforms to the survey conducted by EDUCAUSE [9]. The most common exam proctoring methods used to monitor study participants included: lockdown browser ($n = 71$; 70 %), webcam recording ($n = 65$; 64 %), screen recording ($n = 61$; 60 %), live proctor ($n = 60$; 60 %), and microphone recording ($n = 51$; 50 %) (**Q23**). (See Figure 5.)

While most participants ($n = 94$; 92 %) reported that at least one of their online proctored exams was a course exam (e. g., test, midterm exam, final exam), many ($n = 47$; 46 %) had also used online exam proctoring for lower stakes course assessments such as quizzes (**Q2**). The most common subjects that were proctored included science ($n = 24$; 24 %), business ($n = 17$; 17 %), mathematics ($n = 16$; 16 %), computer science ($n = 11$; 11 %), and medicine ($n = 9$; 9 %).

Many of the participants ($n = 83$; 81 %) took their most recent online proctored exam in the year 2020 during Covid-19, and most were in the last half of 2020: December ($n = 39$; 47 %), November, ($n = 16$; 19 %), and October ($n = 8$; 10 %) (**Q4**). This matches the explosive growth in browser reviews described in Section 3.

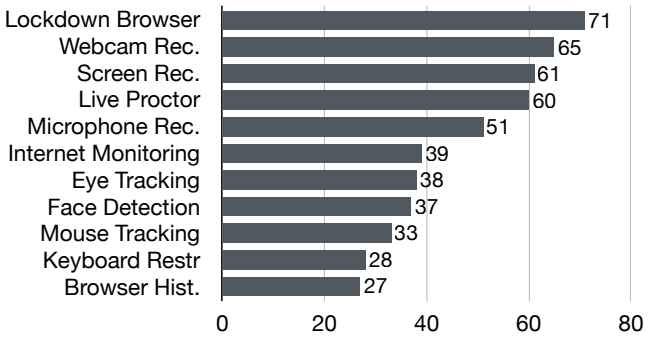


Figure 5: Prevalence of monitoring types (**Q23**).

Requirement to Use Online Proctoring Nearly all participants were required to use an online proctoring service: 97 % of subjects ($n = 99$) noted they had been required to take an exam using online proctoring services (**Q7**) by an authority at their university. When asked who had required them to take an online proctored exam (**Q8**), 70 % ($n = 68$) of respondents indicated their class instructor, followed by 23 % ($n = 22$) who reported that online proctoring was required by their university. Only 7 % ($n = 7$) of participants indicated their requirement to use online proctoring had stemmed from having taken a standardized test.

Preference for/against Online Proctoring A majority of participants ($n = 58$; 56 %) prefer traditional exam formats, but others ($n = 30$; 30 %) preferred online proctored exams (**Q12**; Figure 3). Still, half ($n = 51$; 50 %) stated that they *agree* ($n = 36$; 35 %) or *strongly agree* ($n = 15$; 15 %) that online exam proctoring is a good solution for monitoring remote exams (**Q39**). We asked participants to qualitatively explain some of the benefits of using online exam proctoring (**Q10**): 42 % ($n = 43$) highlighted that they prevent cheating, e.g., "It effectively prevents cheating so students abilities can be graded accurately" (P51); and 29 % ($n = 30$) liked taking exams remotely, e.g., "You don't have to leave your house to take the exam" (P102). Some participants ($n = 5$; 5 %) specifically mentioned the social distancing during the Covid-19 pandemic, e.g., "It was a good way to still be able to take exams securely while distance learning because of COVID-19" (P35). Other participants ($n = 12$; 12 %) liked the flexibility, e.g., "It is a bit nicer to be able to take the exam at a different time that works best for me" (P3).

To explore the factors that may drive a preference for or against taking an online proctored exam, we performed an ordinal logistic regression. For the outcome variable, we used the Likert response to **Q12**, preference for online exam proctoring over traditional exam formats. The factors we considered were participant responses to questions about the number of exams taken, awareness of monitoring methods, concern about the amount of information collected, general privacy perceptions, privacy trade-off, online exams as a good solution, discomfort with monitoring methods, and concern about sharing information. Each of the considered factors was converted to a binary variable, using the appropriate Likert values as bins. Table 6 in Appendix B presents the full regression table.

We find that those that *agree* or *strongly agree* that online proctoring is a good solution for remote examination were 3.66x more likely to have a higher preference for online exams ($b = 1.30, OR = 3.66, p = 0.01$). A lack of privacy concerns also played a role: those that either *disagree* or *strongly disagree* that online proctored exams are privacy invasive were at a significantly increased likelihood of preferring online proctored exams ($b = 2.21, OR = 9.10, p < 0.001$). Surprisingly, if participants *disagree* or *strongly disagree* that they are concerned about the amount of information being collected, they are 5.8x less likely to prefer online exams ($b = -1.76, OR = 0.17, p = 0.03$). At the same time, participants who noted that they are *uncomfortable* or *very uncomfortable* with observation methods during exams were 2.6x less likely to prefer online exams ($b = -0.95, OR = 0.39.p = 0.05$).

The above suggests that while privacy concerns play a role in students' preference for online proctoring, concerns about data collection may not resonate as a privacy concern. Instead, concern about monitoring methods, as we discuss in Section 5.2, appear to be of higher consequence for participants.

Preventing Cheating Online exam proctoring is perceived as a deterrent to cheating. When asked if online exam proctoring makes it less likely for them to cheat, 63 % ($n = 65$) of participants agreed or strongly agreed, while only 26 % ($n = 26$) disagreed or strongly disagreed (**Q29**). However, 60 % ($n = 61$) agreed or strongly agreed that it would still be possible for them to cheat during an online proctored exam, with only 29 % ($n = 29$) who disagreed or strongly disagreed (**Q30**).

When asked to qualitatively explain their belief about the ability to cheat, 21 % ($n = 21$) responded that a second device such as a smartphone could be used, and 13 % ($n = 13$) reported that notes, cheat sheets, or other materials could be used to cheat. Others ($n = 17$; 17 %) explained that it was difficult to cheat. For example, P93 said, "I think that with so many sources being monitored on the student's end, this would make it extremely difficult for them to cheat." Only 2 % of participants ($n = 2$) reported being accused of cheating by the exam proctoring software (**Q32**).

Experiences with Monitoring When asked to described their overall experience being monitored during their exam (**Q25**), some participants ($n = 26$; 25 %) reported that being monitored was a negative experience. For example, P62 responded, "I felt uncomfortable because I do not like being

watched," and P64 stated,

> *... it felt much more stressful than ... taking an exam in a typical proctored environment. I feared that any little movement or sound may trigger the system and flag me for cheating...*

For a minority of participants ($n = 18$; 18 %), being monitored was a positive experience. For instance, P50 stated, "It was pretty good, I stayed focused on the test."

However, other participants ($n = 15$; 15 %) had privacy concerns about being monitored, including P27 who shared, "It does feel uncomfortable to have my person and screen recorded via video, knowing that the recordings are saved for at least some period of time," and P55 who noted, "Its [sic] terribly intrusive and not worth the possibility that students will cheat."

For some participants ($n = 12$; 12 %), being monitored was a distraction or caused increased stress that was detrimental to their exam performance. For example, P66 indicated, "It creates a very stressful environment that prevents me from working to the best of my abilities," and P22 described it as "icky and uncomfortable" and that they felt like they "had to perform in a certain way because I didn't know if someone was watching."

RQ1 Key Findings Many students took an online-proctored exam in the wake of the Covid-19 pandemic, which corresponds to our analysis of browser extension reviews (Section 3). Participants predominantly did not take online proctored exams by choice but were rather required to do so by their instructors. By and large, participants have taken multiple exams with a remote proctor. At the same time, most respondents would prefer a traditional exam even while acknowledging that online exam proctoring is a good solution for remote exams. Those who think online proctoring is a good solution for remote examination as well as those who do not think proctored exams are privacy invasive are more likely to prefer online exam proctoring. We found that concern about data collection matters less than concern about monitoring methods when it comes to privacy and exam preference. Participants also largely believed that exam proctoring deters cheating, but most felt that it was still possible to cheat, particularly using a second device.

5.2 RQ2: Privacy Concerns

We next investigate students' privacy concerns regarding online exam proctoring (**RQ2**).

Comfort with Monitoring Methods When asked about their general comfort level with the methods used to proctor their exam (**Q22**), participants were slightly more comfortable overall (see Figure 6): 45 % ($n = 46$) were either *comfortable* or *very comfortable* with monitoring, while 37 % ($n = 38$) were either *uncomfortable* or *very uncomfortable*. (The remaining participants ($n = 18$; 18 %) were *neither comfortable nor uncomfortable*.)

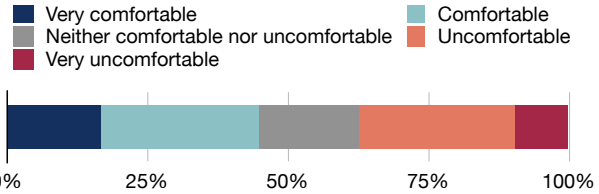


Figure 6: General comfort with proctoring methods (**Q22**).

We also asked participants about their comfort with specific monitoring methods (**Q28**). Aggregated results are presented in Figure 7. To compare the comfort across monitoring methods, we additionally performed a Kruskal-Wallace H-test ($H = 94.6, p < 0.001$) which showed significant difference, and a post-hoc, pair-wise Mann-Whitney U test (with Holm-Sidek correction) indicated that those differences are dominant when comparing monitoring via lockdown browser (participants' most comfortable monitoring method) and all other methods, except for live proctoring and keyboard restrictions. (See Table 7 in Appendix B.) In particular, there are significant differences with some of the most common monitoring methods: webcam recording, screen recordings, and microphone recordings. This suggests that some of the methods deemed most invasive are among those that are used most often, and this in turn may drive students' privacy concerns.

To explore the factors affecting monitoring comfort further, we performed an ordinal logistic regression with an outcome variable of the Likert response to **Q22** (overall comfort with exam privacy) to reported comfort with individual proctoring methods, binning *comfortable* and *very comfortable*. The full regression table (Table 5) appears in Appendix B. We find that comfort with live proctoring ($b = 1.20, OR = 3.31, p < 0.001$) and webcam recordings ($b = 1.96, OR = 7.08, p < 0.001$) significantly increased the likelihood of being more comfortable with exam proctoring generally, suggesting that discomfort with these forms of observation is problematic for many students; both were commonly experienced, cf. Figure 5.

Participants were also asked how necessary a given monitoring method is for online proctoring (see Figure 8). Again there is a significant difference (Kruskal-Wallace: $H = 92.9, p < 0.001$), and a post-hoc analysis (see Table 8 in Appendix B) revealed that there are significant differences between lockdown browser (deemed most necessary) and live proctoring, microphone recording, browser history monitoring, keyboard restrictions, eye tracking and mouse tracking. There were no differences between the trio of lockdown browser, webcam and screen recording with respect to how necessary they are perceived to be for online proctoring. Webcam and screen recording were not (pair-wise) significantly different than live proctoring.

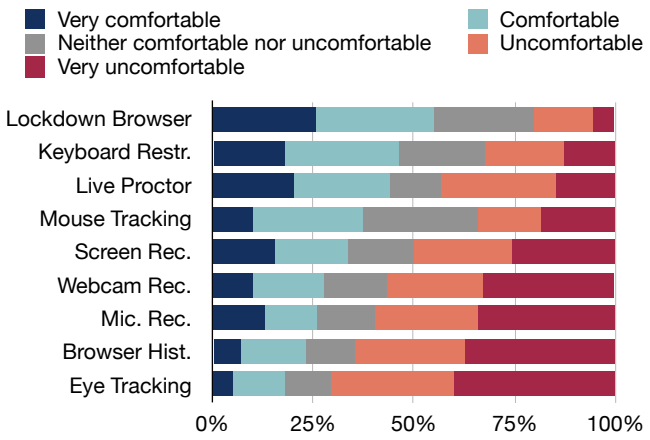


Figure 7: Comfort with monitoring types (**Q28**).

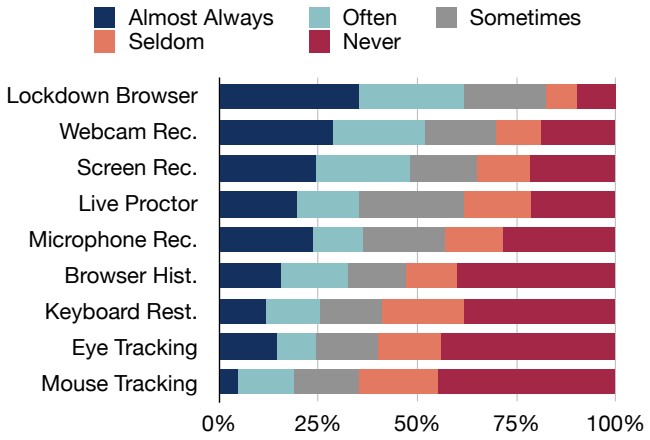


Figure 8: Necessity of monitoring types (**Q27**).

Sharing Information Part of the process of taking an online proctored exam is to verify the identity of the exam taker. This process may involve proof of identification via physical documentation such as IDs and other forms of identity checks that may require students to provide sensitive information to online proctoring services. Students may also be required to create accounts on these services to facilitate that process, and we find that 44 % ($n = 45$) of study participants were required to do just that (**Q18**). Participants also reported that many forms of personal information were required during account creation and before taking an exam, such as full name ($n = 56$; 55 %), student ID number ($n = 52$; 51 %), email address ($n = 51$; 50 %), educational institution ($n = 39$; 38 %), birth date ($n = 29$; 28 %), phone number ($n = 19$; 19 %), residential address ($n = 16$; 16 %), driver's licence number ($n = 10$; 10 %), and social security number ($n = 7$; 7 %) (**Q19**; see Figure 11 in Appendix B). For some participants, physical documentation was required; these included student IDs ($n = 56$; 55 %), driver's licenses ($n = 32$; 31 %), and passports ($n = 7$; 7 %) (**Q20**; see Figure 12 in Appendix B). When asked if they were concerned about sharing this kind of information with online exam proctoring companies, most participants ($n = 62$; 61 %) *agreed* ($n = 38$; 37 %) or *strongly agreed* ($n = 24$; 24 %) (**Q34**). Of those who responded with concerns (**Q35**), being uncomfortable sharing personal information was the most common explanation ($n = 28$; 27 %). For instance, P91 shared, "I feel uneasy that in order to take an exam, I have to share personal information," and P45 said, "I understand that if I opt to take a test online it needs to be fairly taken, but that doesn't mean I should open up these proctoring companies up to my home…"

Data collection was also a concern for some participants ($n = 17$; 17 %). For example P101 responded, "I am not sure what they will do with my information and how long they will store/keep my information," and P58 shared, "For things like recording my computer, or accessing my browser history, I feel like that could invite abuse that go beyond simply making sure I'm honestly taking an exam…" Other participants ($n = 28$; 27 %) had no concerns about sharing information with exam proctoring services, such as P53, who said, "I'm not anymore [sic] worried about it than I am sharing my info with the school," and P48, who said, "I feel since it was required by my school it is a safe place to share information."

Privacy Trade-off When asked if they thought online exam proctoring was too privacy invasive, 52 % ($n = 53$) of study participants *agreed* ($n = 25$; 25 %) or *strongly agreed* ($n = 28$; 27 %) that it was too privacy invasive (**Q36**). There was a split between those who agreed that online exam proctoring offered a reasonable trade-off between personal privacy and exam integrity and those who disagreed (**Q37**). Forty-one percent ($n = 42$) of participants *agreed* ($n = 30$; 29 %) or *strongly agreed* ($n = 21$; 21 %) while 39 % ($n = 39$) of participants *disagreed* ($n = 24$; 24 %) or *strongly disagreed* ($n = 15$; 15 %). We also find evidence of split opinions regarding online proctoring in the qualitative results. In response to **Q11**, $n = 11$ (11 %) of participants reported that there was a trade-off between privacy and academic integrity. For example, P41 noted, "I think it is a valid reason to use online exam proctoring … during this time pandemic. … I can understand giving up some privacy to ensure integrity of exam results."

Participants also reported being concerned about the *amount* of information that online proctoring services collect during the exam (**Q38**). Fifty-seven percent ($n = 58$) of participants *agreed* ($n = 33$; 32 %) or *strongly agreed* ($n = 25$; 25 %), while 26 % ($n = 26$) of participants *disagreed* ($n = 21$; 21 %) or *strongly disagreed* ($n = 5$; 5 %). We again see similar results in qualitative responses in **Q11**: 59 % ($n = 60$) reported a privacy concern, with concerns about webcam access being the most common ($n = 27$; 26 %). For example, P65 reported, "I believe that online exams can be invasive, as at least mine required both a webcam and microphone, so they could see me and my room and hear my surroundings," and P36 responded:

> *. . . Unlike in-class proctoring, students must be filmed in their homes . . . The view is also on the student* 100 % *of the time so the student cannot relax and has their entire body language and quirks on display. It is a breach of privacy without enough benefit to justify it.*

Some participants ($n = 6$; 6 %) had concerns about relinquishing control of their computing devices to the exam proctoring services, e.g., "It is a little scary about how much they can access and control your device" (P101). Sharing of personal information was a concern for participants ($n = 6$; 6 %), such as P39, who noted, "It does make me a little uncomfortable that there is a 3rd party company that may have my personal identification and see into my room." Still other participants ($n = 19$; 19 %) reported that they had no privacy concerns, such as P69, who said, "I don't see any huge issues with privacy in online exam proctoring," and P51 who stated:

> *I don't mind that they can see my room and control my screen. They aren't doing anything sinister, and I can revoke all permissions at the end of the exam.*

RQ2 Key Findings A majority of students found online exam proctoring to be privacy invasive, most citing concerns with the webcam and microphone recordings, which provides the means to view, listen, and record inside a student's room. However, some students felt a trade-off between loss of personal privacy and exam integrity was reasonable.

When considering privacy in the context of preventing academic dishonesty, participants had mixed reactions to the proctoring methods used. Lockdown browsers, webcams, and screen recordings were viewed as necessary for online exam proctoring compared to other methods, and these were also the most commonly used methods to observe students during an exam. However, only about a quarter of the respondents were comfortable with webcam or screen recording, while half were comfortable with lockdown browsers. Regression analysis indicated that comfort with live proctoring and webcam recordings drive comfort overall with monitoring. This suggests that there is a gap between the proctoring methods commonly used and the comfort level of students with those observation techniques.

Students also create accounts and share significant information with online proctoring services to verify their identities. Services often require personal information, such as a student ID number, email address, phone number, and residential address, as well as images of physical documentation such as student IDs and driver's licenses. Participants expressed concern about sharing this and other personal information with exam proctoring companies. They were also concerned about the overall quantity of information collected, what would be done with the information, and how long it would be stored.

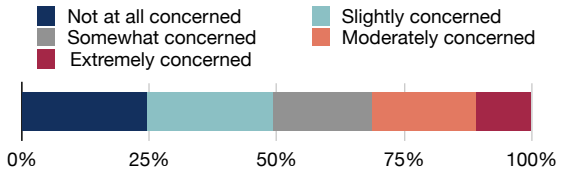


Figure 9: Concern over installing proctoring software (**Q49**).

5.3 RQ3: Security Concerns

Along with privacy concerns, the installation of specialized software to enable proctoring could lead to security issues, and as part of addressing **RQ3**, we surveyed participants about their experiences and concerns with proctoring software.

Browser Extensions One way in which exam proctoring companies provide monitoring during an exam is by requiring students to install web browser extensions, as noted in Section 3. Most study participants ($n = 65$; 64 %) were required to install a web browser extension to take their exam (**Q40**). When we asked participants who installed a browser extension what they thought the extension did (**Q41**), they responded that the extension locked down their browser ($n = 27$; 42 %), collected data ($n = 8$; 12 %), monitored network activity ($n = 8$; 12 %), initiated screen recording ($n = 7$; 11 %), disabled functionality on their device ($n = 6$; 9 %), and enabled their webcam ($n = 6$; 9 %) and microphone ($n = 4$; 6 %).

Proctorio ($n = 13$; 13 %) was the most common web browser extension installed by study participants, followed by ProctorU ($n = 12$; 12 %), and Honorlock ($n = 8$; 8 %) (**Q42**; Figure 13 in Appendix B). Significantly, only 45 % ($n = 31$) of participants who installed a web browser extension reported removing or disabling the extension after completing their exam (**Q43**). Given that many of these browser extensions have pervasive monitoring permissions that can be activated on a broad set of URLs (see Section 3), it is important that students remove these extensions; the failure of 45 % ($n = 30$) to do so suggests that installing this custom software may put students at risk beyond the exam.

Standalone Software Another way in which exam proctoring companies provide monitoring during the examination is by requiring students to install standalone software, which we define as software that is installed as an application on their computer and is not a browser extension. Thirty-five percent ($n = 36$) of participants reported they were required to install exam proctoring software (not including a browser extension) (**Q44**). When we asked participants who installed exam proctoring software what they thought the software did (**Q45**), they responded that the proctoring software locked down their browser ($n = 11$; 31 %), disabled functionality on their device ($n = 8$; 22 %), monitored their activity ($n = 7$; 19 %),

initiated screen recording ($n = 3$; 8 %), and enabled their webcam ($n = 3$; 8 %). Of the participants who installed exam software, most ($n = 32$; 89 %) said that they did uninstall the exam proctoring software after the exam was complete (**Q44**). Only one participant reported having issues uninstalling the exam proctoring software (**Q44**). Most participants ($n = 88$; 86 %) said they had installed the exam software on their personal computer (**Q44**). When asked to report their concern for installing proctoring software (Figure 9), 52 % of participants ($n = 52$) specified that they were at least *somewhat concerned* ($n = 20$; 20 %), *moderately concerned* ($n = 21$; 21 %), or even *extremely concerned* ($n = 11$; 11 %). In contrast, 48 % ($n = 50$) of participants were *slightly concerned* ($n = 25$; 24 %) or *not at all concerned* ($n = 25$; 24 %) (**Q49**).

We asked participants to explain their concern, or lack of concern, regarding the installation of online exam software (**Q49**). Many participants ($n = 49$; 48 %) explained that they had privacy concerns. Some ($n = 11$; 11 %) replied that the software's potential access to sensitive personal information was a concern; for instance, P32 said, "I am worried it will be able to access sensitive information," and P38 reported, "It is my own computer which stores all of my information so that is a bit iffy." A few participants ($n = 6$; 6 %) had concerns about data collection from their computer after the exam was completed, such as P52, who said, "I wonder if they continued collecting information after the exam," and P102, who noted, "I don't trust software that is designed to gather information about my activities to confine itself to being used only for exams." Others ($n = 11$; 11 %) were unsure what information could be collected by the software; e.g., P69 and P34, who respectively stated "It's unclear what all data it's collecting and when it's running," and "I don't know what information it was collecting or how it would be used." Still others ($n = 28$; 27 %) had no concerns about the software. A few ($n = 5$; 5 %) stated they had no concerns because the exam privacy software was supported by their university. For instance, P40 noted, "I believe the university would not use the proctoring service if their software was dangerous," and P87 stated, "I know that my school and professors wouldn't have me install anything that could harm my computer or invade my privacy."

RQ3 Key Findings The browser extension is the most common way in which exam proctoring tools access students computing devices. Students understand that these extensions are used both to surveil them and their devices during the exam and to disable functionality that would otherwise allow them to access unauthorized resources. Despite this knowledge, only a small number of students actually removed or disabled the extension after completing their exam, leaving permission-hungry software residing on their computers.

Standalone software is also used for exam proctoring, and most students install this software on their own personal computers. Students are concerned about this software and say they worry that it may access personal information stored on their computers. Most students did uninstall this standalone software, an action that highlights and confirms their concerns.

6 Recommendations and Conclusions

Privacy Trade-offs During a Pandemic Given the necessary rapid transition to remote learning, many institutions and educators did not have sufficient time to restructure courses around alternative forms of learning and skills assessment. Content and exams that had originally been envisioned as in-person suddenly had to be delivered and proctored remotely, forcing institutions to seek solutions to a perceived exam integrity problem. Our results suggest that students understand that their educational institutions were struggling to maintain mandated safety protocols while continuing to provide academic rigor, as evidenced by the fact that a large number of study respondents (41 %) reported that online exam proctoring was a reasonable trade-off between personal privacy and exam integrity. A recurring theme in the qualitative responses was that giving up some personal privacy during the Covid-19 pandemic to maintain safety protocols was a valid reason to accept the use of online proctoring services. This suggests that student acceptance of online exam monitoring is higher than it might otherwise be in a post-pandemic situation.

At the same time, we find that a large percentage of students have significant concerns—e.g., sharing personal information with proctoring companies, the amount of information collected by these companies, and installing online exam proctoring software on their computers. When we consider these facts, it is clear that many students found their proctored exams to be privacy invasive and would prefer alternatives to online proctored exams. However, it is unclear if a post-pandemic context will lead to increased student opposition to invasive monitoring or if students will have become accustomed to these proctoring tools and resigned to their use.

Recommendation: Based on this study, we recommend that institutions and instructors both expand student choice by developing alternative forms of student assessment that can account for privacy concerns whenever practicable and plan to reduce future reliance on exam proctoring services after the Covid-19 pandemic.

Necessary Type of Monitoring Many participants agreed that the ability to deter cheating during an exam is important, up to a point, after which they felt that monitoring goes from necessary to unnecessary and invasive. In fact, we find that the types of monitoring that students perceive as the most unnecessary are among those that they report are the most uncomfortable and invasive. For instance, a majority of students reported that they do not think it is necessary to monitor mouse movement, eye movement, or web browser history; correspondingly, a majority also reported being uncomfort-

able with the monitoring of their eye movement, web browser history, microphone, and webcam. These very monitoring types are those that students refer to most when they discuss how they feel that the online proctored exam can create a stressful environment, how they feel "watched," and how they worry that any small sound or tiny movement—such as looking away from the screen briefly—could flag them for cheating. Students report that these additional stressors and anxieties distract them, reduce their focus, and prevent them from performing to the best of their abilities. This level of monitoring assumes cheating and pre-penalizes all students with additional stress and anxiety whether they were planning to be honest or dishonest.

Our work suggests that even though technologically advanced invigilation techniques are available, such as 360-degree room scans and eye movement tracking, it does not mean that they are either necessary to curb cheating or sensitive to students' personal privacy and device security. Continuing to use monitoring techniques that students find unnecessary for exam integrity, while at the same time requiring students to sacrifice their personal privacy, displays a lack of trust for students and undermines students' trust in educators and institutions.

Recommendation: We recommend that institutions and educators follow a principle of least monitoring by using the minimum number of monitoring types necessary, given the class size and knowledge of expected student behavior. Institutions should perform due diligence when selecting online exam proctoring companies with whom to contract, and they should take into account student privacy, student discomfort for certain monitoring types, and software installation requirements. Moreover, instructors should use caution when selecting monitoring types while setting up exams and should provide students with clear reasoning for having selected the individual methods that will be used to monitor their exams.

Invasion of Personal Computers As we have seen, exam proctoring browser extensions and standalone software contain invasive monitoring tools. These tools often include permissions to access the webcam, microphone, and web browser history. However, 43 % of students did not remove or disable the required browser extensions once they had completed their assessments. As is the case with any custom software, there is risk of vulnerabilities in these extensions. The fact that students often neglect to remove them therefore creates increased potential for harm from loss of privacy or security intrusions. Institutions should therefore be sensitive to which proctoring software they require students to install on their personal devices.

Recommendation: We recommend that institutions thoroughly review common vulnerabilities and exposures of the online exam proctoring software they plan to license for installation on students' personal computers. We would also recommend that institutions limit the installation of standalone exam proctoring software to devices issued to students by the institution.

Implied Trust via Institutional Support Exam proctoring tools are often integrated with existing learning management software, such as Blackboard and Canvas, giving the appearance that they are a part of the standard educational software stack and imparting a sense of safety and normalcy. At the same time, institutions have spent large amounts of money to obtain site license agreements for exam proctoring software. This gives the appearance of a certain amount of due diligence being applied to the purchase, while potentially increasing the barriers to student resistance towards these forms of examination.

Throughout our qualitative findings are statements from students of a transfer of trust between institutions who licence and faculty who support the exam proctoring software and the software itself. These students say that they believe their university would not use the software to proctor exams if it was dangerous. Moreover, they believe that their school would not have them install anything that could harm their computer or invade their privacy. Institutional support for third-party proctoring software, which conveys credibility, makes the exam proctoring software appear safer and less potentially problematic because students assume that institutions have done proper vetting of both the software and the methods employed by the proctoring services.

Recommendation: We recommend that the students, along with faculty and administrators, take part in the assessment and selection of exam proctoring software. Students should be involved in every step of the process, from deciding whether to use exam proctoring software to determining which, if any, software should be used and which methods should be made available for exam monitoring.

Power Imbalances Finally, when students' options are limited to taking an online proctored exam or failing the course, it is a clear indication of an institutional power dynamic; 97 % of students indicated they were required to take an online proctored exam. Offering students more choices for assessment and being upfront with students about institutional privacy norms is a crucial step to alleviate this power imbalance.

Recommendation: We recommend implementing notice and choice for courses employing online exam proctoring and allowing students to consent to any monitoring that will take place during course quizzes and exams. We also recommend that syllabi include a readable privacy policy to better communicate expectations.

Acknowledgements

This work is partially funded by the National Science Foundation under grants 1718498 and 1845300, and the Georgetown University Callahan Family Professor of Computer Science Chair Fund.

References

[1] Olivia Buccieri. Online Exam Proctoring No Longer Allowed for UC Berkeley Classes. *The Daily Californian*. https://www.dailycal.org/2020/04/05/online-exam-proctoring-no-longer-allowed-for-uc-berkeley-classes/, April 2020.

[2] Lilah Burke. Cutting the in-person semester short. *Inside Higher Ed*. https://www.insidehighered.com/news/2020/11/17/colleges-end-person-instruction-early-due-covid-19-spread, November 2020.

[3] Simon Coghlan, Tim Miller, and Jeannie Paterson. Good proctor or "Big Brother"? AI Ethics and Online Exam Supervision Technologies. *arXiv preprint arXiv:2011.07647*, 2020.

[4] Shaanan Cohney, Ross Teixeira, Anne Kohlbrenner, Arvind Narayanan, Mihir Kshirsagar, Yan Shvartzshnaider, and Madelyn Sanfilippo. Virtual Classrooms and Real Harms. *arXiv preprint arXiv:2012.05867*, 2020.

[5] ConductExam. Create, share & analyze exams with the best online exam software. https://www.conductexam.com.

[6] Lee William Daffin Jr and Ashley A Jones. Comparing Student Performance on Proctored and Non-Proctored Exams in Online Psychology Courses. *Online Learning*, 22(1):131–145, 2018.

[7] Colleen Flaherty. Big proctor. *Inside Higher Ed*. https://www.insidehighered.com/news/2020/05/11/online-proctoring-surging-during-covid-19, May 2020.

[8] Patricia A Goedl and Ganesh B Malla. A Study of Grade Equivalency between Proctored and Unproctored Exams in Distance Education. *American Journal of Distance Education*, pages 1–10, 2020.

[9] Susan Grajek. EDUCAUSE COVID-19 QuickPoll Results: Grading and Proctoring. EDUCAUSE Research Notes, April 2020.

[10] Therese C Grijalva, Joe Kerkvliet, and Clifford Nowell. Academic Honesty and Online Courses. *College Student Journal*, 40(1), 2006.

[11] Drew Harwell. Mass School Closures in the Wake of the Coronavirus are Driving a New Wave of Student Surveillance. *The Washington Post*. https://www.washingtonpost.com/technology/2020/04/01/online-proctoring-college-exams-coronavirus/, April 2020.

[12] Honorlock. Online Exam Proctoring with a Human Touch. https://honorlock.com.

[13] Shawn Hubler. Keeping Online Testing Honest? Or an Orwellian Overreach? *The New York Times*. https://www.nytimes.com/2020/05/10/us/online-testing-cheating-universities-coronavirus.html, May 2020.

[14] Kenrie Hylton, Yair Levy, and Laurie P Dringus. Utilizing Webcam-based Proctoring to Deter Misconduct in Online Exams. *Computers & Education*, 92:53–63, 2016.

[15] IRIS Invigilation. Assess with integrity. https://www.irisinvigilation.com.

[16] Faten F Kharbat and Ajayeb S Abu Daabes. E-proctored Exams During the COVID-19 Pandemic: A Close Understanding. *Education and Information Technologies*, pages 1–17, 2021.

[17] Allison Kushner and Kevin Pitts. Letter to instructors, 2021. University of Illinois, Urbana-Champaign.

[18] Mark M Lanier. Academic Integrity and Distance Learning. *Journal of criminal justice education*, 17(2):244–261, 2006.

[19] Doug Lederman and Mark Lieberman. How Many Public Universities Can 'Go Big' Online? *Inside Higher Ed*. https://www.insidehighered.com/digital-learning/article/2019/03/20/states-and-university-systems-are-planning-major-online, March 2019.

[20] Naresh K. Malhotra, Sung S. Kim, and James Agarwal. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Info. Sys. Research*, 15(4):336–355, December 2004.

[21] Mercer Mettl. Conduct Extremely Secure, Scalable, And Cost-Effective Online Proctored Exams With One Click. https://mettl.com.

[22] Derek Newton. Another Problem with Shifting Education Online: Cheating. *Hechinger Report*. https://hechingerreport.org/another-problem-with-shifting-education-online-cheating/, August 2020.

[23] Anushka Patil and Jonah Engel Bromwich. How It Feels When Software Watches You Take Tests. *The New York Times*. https://www.nytimes.com/2020/09/29/style/testing-schools-proctorio.html, September 2020.

[24] ProctorExam. Leading online proctoring services. https://proctorexam.com.

[25] Proctorio. A Comprehensive Learning Integrity Platform. https://proctorio.com.

[26] ProctorU. Online Proctoring to Advance your Learning and Testing Program. https://www.proctoru.com.

[27] PSI Online. Where People Meet Potential. https://www.psionline.com.

[28] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples. In *IEEE Symposium on Security and Privacy (SP)*, May 2019.

[29] Respondus. Assessment Tools for Learning Systems. https://web.respondus.com.

[30] Joseph A. Rios and Ou Lydia Liu. Online Proctored Versus Unproctored Low-Stakes Internet Test Administration: Is There Differential Test-Taking Behavior and Performance? *American Journal of Distance Education*, 31(4):226–241, 2017.

[31] Shea Swauger. Software that Monitors Students During Tests Perpetuates Inequality and Violates their Privacy. *MIT Technology Review*, August 2020.

[32] George R Watson and James Sottile. Cheating in the Digital Age: Do Students Cheat More in Online Courses? In *Society for Information Technology & Teacher Education International Conference*. Association for the Advancement of Computing in Education (AACE), 2008.

[33] Daniel Woldeab and Thomas Brothen. 21st Century Assessment: Online Proctoring, Test Anxiety, and Student Performance. *International Journal of E-Learning & Distance Education*, 34(1), 2019.

A Survey Instruments

A.1 Screening Survey

S1 How familiar are you with online exam proctoring?
- ◯ Not at all familiar
- ◯ Slightly familiar
- ◯ Somewhat familiar
- ◯ Moderately familiar
- ◯ Extremely familiar

S2 I have taken an online proctored exam.
- ◯ Yes
- ◯ No
- ◯ Unsure
- ◯ Prefer not to answer

S3 Please describe your overall online proctored exam experience.

Answer: _____

These questions were followed by the 10 IUIPC items as described by Malhotra et al. [20]

D1 What is your gender?
- ◯ Woman
- ◯ Man
- ◯ Non-binary
- ◯ Prefer not to disclose
- ◯ Prefer to self-describe

D2 What is your age?
- ◯ 18 – 24
- ◯ 25 – 34
- ◯ 35 – 44
- ◯ 45 – 54
- ◯ 55 – 64
- ◯ 65 or older
- ◯ Prefer not to disclose

D3 Are you currently a student?
- ◯ Yes
- ◯ No
- ◯ Prefer not to disclose

D4 What is the highest degree or level of school you have completed?
- ◯ No schooling completed
- ◯ Some high school, no diploma
- ◯ High school graduate, diploma, or equivalent (e. g., GED, Abitur, baccalaureat)
- ◯ Some college credit, no degree
- ◯ Trade / technical / vocational training
- ◯ Associate degree
- ◯ Bachelor's degree
- ◯ Master's degree
- ◯ Professional degree (e. g., J.D., M.D.)
- ◯ Doctorate degree
- ◯ Prefer not to disclose

D5 Which of the following best describes your educational background or job field?
- ◯ I have an education in, or work in, the field of computer science, computer engineering or IT.
- ◯ I do not have an education in, nor do I work in, the field of computer science, computer engineering or IT.
- ◯ Prefer not to disclose

A.2 Main Study

Thank you for participating in the second part of our survey. You have been invited to our main study because of your direct experience taking an online proctored exam.

Your answers based on your online exam proctoring experiences are important to us!

Please read the following instructions carefully:
- Take your time in reading and answering the questions.
- Answer the questions as accurately as possible.
- It is okay to say that you don't know an answer.

Q1 How many online proctored exams have you taken?
- ◯ 1
- ◯ 2
- ◯ 3
- ◯ 4
- ◯ 5+

Q2 What was the nature of the exam(s) you took using an online proctoring service? Select all that apply.
- ◯ Course Quiz
- ◯ Course Exam (E.g. test, midterm exam, final exam)
- ◯ Standardized Test (E.g. GRE, GMAT, bar exam)
- ◯ I have not taken an exam with online proctoring
- ◯ Other: _____

Q3 Of those ones you chose, which is the most recent?
- ◯ Course Quiz
- ◯ Course Exam (E.g. test, midterm exam, final exam)
- ◯ Standardized Test (E.g. GRE, GMAT, bar exam)
- ◯ I have not taken an exam with online proctoring
- ◯ *[Other value entered in **Q2**]*

Q4 As best as you can remember, when was the month, date, and year when you last took an online exam with a proctoring service?
- ◯ Month: _____
- ◯ Day: _____
- ◯ Year: _____

Q5 What was the subject matter of the last examination you took using an online proctoring service?

Answer: _____

Q6 What was the name of the online proctoring service used during your last examination?
- ◯ ConductExam
- ◯ Pearson OnVUE
- ◯ PSI Online Proctoring
- ◯ Examity
- ◯ ProctorExam
- ◯ Questionmark
- ◯ ExamSoft
- ◯ ProctorFree
- ◯ Respondus
- ◯ Honorlock
- ◯ Proctorio
- ◯ Smowl
- ◯ IRIS Invigilation
- ◯ Proctortrack
- ◯ Surpass
- ◯ Kryterion
- ◯ ProctorU
- ◯ Talview
- ◯ Mercer Mettl
- ◯ Proview
- ◯ TestReach
- ◯ Other: _____
- ◯ Unsure

Q7 Were you required to take that exam using an online exam proctoring service?
- ◯ Yes, I was required to use an online exam proctoring service.
- ◯ No, there were other forms of assessment available to me but I opted to use an online exam proctoring service

Q8 Who required you to take an online proctored exam?

Answer: _____

Q9 What were the deciding factors in your choice to take your exam with an online proctoring service instead of other forms of assessment?

Answer: _____

Q10 In your experience, what are some benefits of using online exam proctoring?

Answer: _____

Q11 Please explain your views on the privacy of online exam proctoring.

Answer: _____

Q12 I prefer online exam proctoring services over traditional exam formats.
- ◯ Strongly disagree
- ◯ Disagree
- ◯ Neither agree nor disagree
- ◯ Agree
- ◯ Strongly agree

Q13 Did the online proctoring service make any necessary exam accommodations or other modifications based on your needs as an exam taker?
- ◯ Yes, I request and was provided adequate accommodations
- ◯ No, I requested and was not provided adequate accommodations
- ◯ I did not request nor require exam accommodations
- ◯ Unsure
- ◯ Prefer not to answer

Q14 Please describe the accommodations provided to you. You may indicate N/A if you prefer not to answer. *[Shown only if answer to **Q13** was "Yes"]*

Answer: _____

Q15 Please describe how accommodations were not provided for you despite your request. You may indicate N/A if your prefer not to answer. *[Shown only if answer to **Q13** was "No"]*

Answer: _____

Q16 Did you experience any technical difficulties when taking your exam as it relates to the online proctored service?

◯ Yes ◯ No ◯ Unsure

Q17 Please explain any technical difficulties you may have experienced during your exam with the online proctored service. *[Shown only if answer to **Q15** was "Yes"]*

Answer: _____

Online Exam Proctoring Methods

In this part of the survey you will be asked about the methods employed by online exam proctoring services.

Q18 When preparing to take an online proctored exam were you required to create an account with the online proctoring service?

◯ Yes ◯ No ◯ Unsure

Q19 When registering for an online proctored exam what, if any, personal information where you required to enter in online forms? Select all that apply.

◯ Residential Address
◯ Educational institution affiliation
◯ Email Address
◯ Student ID Number
◯ Full Name
◯ Social Security Number
◯ Driver's License Number
◯ Phone Number
◯ No information was required
◯ Unsure
◯ Other: _____

Q20 When taking an online proctored exam what kinds of physical documentation, if any, were you required to provide? Select all that apply.

◯ Driver's License
◯ Student ID
◯ Passport
◯ No physical documentation was required
◯ Unsure
◯ Other: _____

Q21 How aware are you of the methods used by online exam proctoring services to monitor exam takers?

◯ Not at all aware
◯ Slightly aware
◯ Somewhat aware
◯ Moderately aware
◯ Extremely aware

Q22 How comfortable were you with the methods used to proctor the exam(s) that was proctored online?

◯ Very comfortable
◯ Comfortable
◯ Neither comfortable nor uncomfortable
◯ Uncomfortable
◯ Very uncomfortable

Q23 Please select all methods that were used to proctor the exam(s) that was proctored online. Select all that apply.

◯ Live proctor visible to me
◯ Live proctor not visible to me
◯ Web browser history monitoring
◯ Eye movement tracking
◯ Facial detection
◯ Lockdown browser
◯ Mouse movement tracking
◯ Keyboard restrictions (E.g. no copy and paste)
◯ Screen recording
◯ Microphone recording
◯ Internet activity monitoring (E.g. interaction with a web site)
◯ Webcam recording

Q24 Please describe any methods not listed above, or indicate "none" if there are no more methods.

Answer: _____

Q25 Please describe your overall experience being monitored during your online proctored exam.

Answer: _____

Q26 Previously you indicated that your most recent online proctored exam was a *[Answer from **Q3**]*. Please refer to that experience in answering the following questions. Considering your most recent online proctored exam, what kind of monitoring (if any) was necessary to proctor that exam online?

Answer: _____

Q27 Previously you indicated that your most recent online proctored exam was a *[Answer from **Q3**]*. Again considering your most recent exam from the question above. For each exam monitoring type please select how often they are necessary for online proctoring.

	Always	Often	Rarely Sometimes		Never
Live proctor	◯	◯	◯	◯	◯
Web browser history monitoring	◯	◯	◯	◯	◯
Eye movement tracking	◯	◯	◯	◯	◯
Lockdown browser	◯	◯	◯	◯	◯
Mouse movement tracking	◯	◯	◯	◯	◯
Keyboard restrictions (E.g. no copy/paste)	◯	◯	◯	◯	◯
Screen recording	◯	◯	◯	◯	◯
Microphone recording	◯	◯	◯	◯	◯
Webcam recording	◯	◯	◯	◯	◯

Q28 Again considering your most recent online proctored exam of *[Answer from **Q3**]*. For each exam monitoring type please select how comfortable you feel about them.

	Very Comfortable	Comfortable	Neither comfortable nor uncomfortable	Uncomfortable	Very Uncomfortable
*[Types from **Q27**]*	◯	◯	◯	◯	◯

Online Exam Proctoring Functionality

In this part of the survey you will be asked about the functionality of online exam proctoring services. If you feel uncomfortable answering any question below, you may select "Prefer not to answer."

Q29 The use of online exam proctoring tools makes it less likely that my classmates or I will cheat on an exam.

◯ Strongly disagree
◯ Disagree
◯ Neither agree nor disagree
◯ Agree
◯ Strongly agree
◯ Prefer not to answer

Q30 If I wanted to, I believe I would still be able to cheat even with online exam proctoring.

◯ Strongly disagree
◯ Disagree
◯ Neither agree nor disagree
◯ Agree
◯ Strongly agree
◯ Prefer not to answer

Q31 Please explain your belief about the ability to cheat (or not cheat) with online proctored exams. Or write N/A if you prefer not to answer.

Answer: _____

Q32 Have you been accused of cheating by exam proctoring software?

◯ Yes ◯ Prefer not to answer
◯ No

Q33 Which of the following methods employed by the online exam proctoring software was used to accuse you of cheating? Select all that apply. *[Shown only if answer to **Q32** was "Yes"]*

◯ Live proctor visible to me
◯ Live proctor not visible to me
◯ Web browser history monitoring
◯ Eye movement tracking
◯ Facial detection
◯ Lockdown browser
◯ Mouse movement tracking
◯ Keyboard restrictions (E.g. no copy and paste)
◯ Screen recording
◯ Microphone recording
◯ Internet activity monitoring (E.g. interaction with a web site)
◯ Webcam recording
◯ Unsure
◯ Other: _____
◯ Prefer not to answer

Privacy Concerns In this part of the survey you will be asked about the benefits and potential risks you associate with online exam proctoring.

Q34 I am concerned about sharing information with online exam proctoring companies.

◯ Strongly disagree ◯ Agree
◯ Disagree ◯ Strongly agree
◯ Neither agree nor disagree

Q35 Please explain your answer to the previous question regarding the consequences of sharing information.

Answer: _____

Q36 I think online exam proctoring services are too privacy invasive.

◯ Strongly disagree ◯ Agree
◯ Disagree ◯ Strongly agree
◯ Neither agree nor disagree

Q37 I think online exam proctoring offers a reasonable tradeoff between my privacy and the integrity of the exam.

◯ Strongly disagree ◯ Agree
◯ Disagree ◯ Strongly agree
◯ Neither agree nor disagree

Q38 I am concerned about the amount of information that online proctoring services collect during the exam.

◯ Strongly disagree ◯ Agree
◯ Disagree ◯ Strongly agree
◯ Neither agree nor disagree

Q39 I think online exam proctoring is a good solution for monitoring remote examinations.

◯ Strongly disagree ◯ Agree
◯ Disagree ◯ Strongly agree
◯ Neither agree nor disagree

Exam Proctoring Web Browser Extensions A web browser extension is a small software module that is used to extend the functionality of your web browser with additional features. Some online proctoring services require exam takers to install a web browser extension in order to take an exam. In this part of the survey you will be asked questions about your experience using web browser extensions to take online proctored exams.

Q40 Were you required to install and use a web browser extension in order to participate in a proctored online exam?

◯ Yes ◯ No ◯ Unsure

Q41 What do you think the browser extension did? *[Shown only if answer to **Q40** was "Yes"]*

Answer: _____

Q42 What was the most recent web browser extension you installed in order to participate in a proctored online exam? *[Shown only if answer to **Q40** was "Yes"]*

◯ ConductExam ◯ ProctorExam ◯ Unsure
◯ Examity ◯ Proctorio ◯ No browser extension was installed
◯ Honorlock ◯ ProctorU
◯ IRIS Invigilation ◯ PSI Online Proctoring ◯ Other: _____
◯ Mercer Mettl

Q43 Did you remove or disable any browser extensions that you were required to install to take an online proctored exam? *[Shown only if answer to **Q40** was "Yes"]*

◯ Yes ◯ No ◯ Unsure

Exam Proctoring Software Some online proctoring services require exam takers to install standalone application software on a computer, like your PC or Mac, in order to take an exam. In this part of the survey you will be asked questions about your experience installing and using exam application software to take online proctored exams. Please note this may be in addition to the requirement to install a browser extension.

Q44 Did you have to install other types of exam proctoring software (not including a browser extension)?

◯ Yes ◯ No ◯ Unsure

Q45 What do you think this exam proctoring software did? *[Shown only if answer to **Q44** was "Yes"]*

Answer: _____

Q46 Did you uninstall the exam proctoring software? *[Shown only if answer to **Q44** was "Yes"]*

◯ Yes ◯ No ◯ Unsure

Q47 Did you have any issues uninstalling the exam proctoring software? *[Shown only if answer to **Q46** was "Yes"]*

◯ Yes ◯ No ◯ Unsure

Q48 From the computing devices listed below, please select the device you used to take your most recent online proctored exam.

◯ Personal Computer ◯ Mobile Device (Smartphone, Tablet, etc)
◯ Shared Home Computer
◯ School Issued Computer ◯ Unsure
◯ Public Computer (E.g. Library) ◯ Other: _____

Q49 How concerned are you about installing online exam proctoring software on the computer you used to take the exam?

◯ Not at all concerned ◯ Moderately concerned
◯ Slightly concerned ◯ Extremely concerned
◯ Somewhat concerned

Q50 Please explain your answer to the previous question regarding your concern about installing online exam proctoring software.

Answer: _____

B Additional Figures and Tables

Metric	Reddit sample	Prolific sample
Total Participants	27	75
Gender: Man	10 (37%)	42 (56%)
Gender: Woman	14 (52%)	33 (44%)
Gender: Nonbinary	2 (7.4%)	0
Gender: Prefer not to disclose	1 (3.7%)	0
Age: 18-24	17 (63%)	56 (75%)
Age: 25-34	8 (30%)	14 (19%)
Age: 35-44	2 (7.4%)	3 (4.0%)
Age: 45-54	0	2 (2.7%)
Age: 55-64	0	0
Student	21 (78%)	69 (92%)
NonStudent	6 (22%)	6 (8.0%)
Some high school (no diploma)	0	1 (1.3%)
High school graduate, diploma, or equivalent	2 (7.4%)	9 (12%)
Trade / technical / vocational training	1 (3.7%)	0
Some college credit, no degree	5 (19%)	35 (47%)
Associate's degree	3 (11%)	11 (15%)
Bachelor's degree	14 (52%)	15 (20%)
Master's degree	1 (3.7%)	2 (2.7%)
Professional degree (e.g., J.D., M.D.)	1 (3.7%)	1 (1.3%)
Schooling : Other (including PhD)	0	1 (1.3%)
IT background	11 (41%)	20 (27%)
No IT background	15 (56%)	55 (73%)
IT background: Prefer not to disclose	1 (3.7%)	0

Table 4: Participant demographics for Reddit and Prolific participants. Prolific demographics exclude tallies from respondents who only completed the pre-survey.

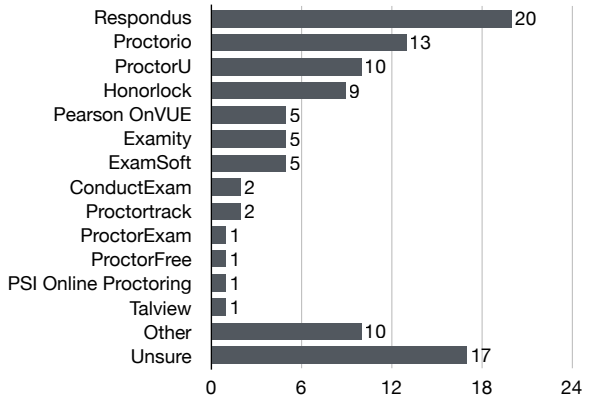


Figure 10: Proctoring services experienced. This generally conforms to the survey conducted by EDUCAUSE [9].

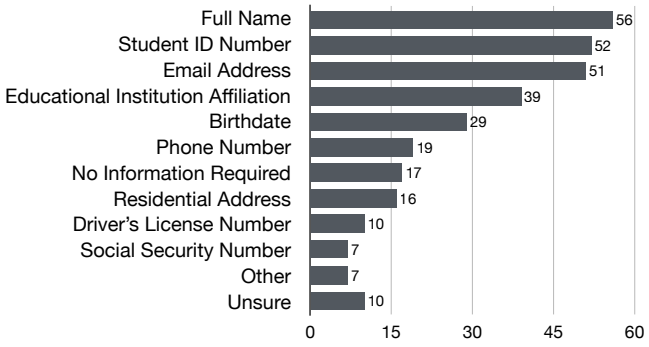


Figure 11: Information required when registering for an online proctored exam (**Q19**).

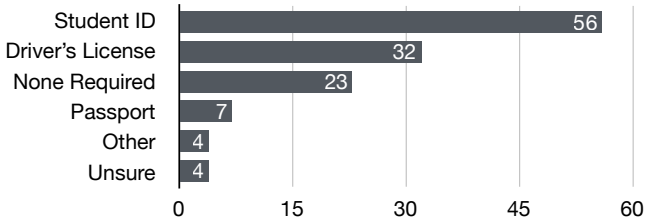


Figure 12: Physical documentation required to provide (**Q20**).

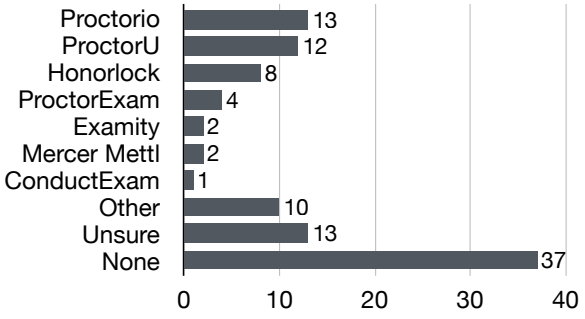


Figure 13: Most common browser extensions installed (**Q42**).

Table 5: Ordinal regression model to describe the level of comfort with proctoring methods responses to Question **Q22**. The model uses an ascending comfort scale (i. e., from *Very uncomfortable* to *Very comfortable*). The Aldrich-Nelson pseudo R^2 of the model is 0.58.

Factor	**Estimate**	**Odds ratio**	**Error**	**t value**	**Pr(>|z|)**	
Live Proctor $\in$ {*Comfortable, Very Comfortable*}	1.20	3.31	0.42	2.88	<0.001	**
Browser History $\in$ {*Comfortable, Very Comfortable*}	−0.49	0.61	0.54	−0.91	0.36	
Eye Tracking $\in$ {*Comfortable, Very Comfortable*}	−0.80	0.45	0.72	−1.11	0.27	
Lockdown Browser $\in$ {*Comfortable, Very Comfortable*}	−0.05	0.95	0.46	−0.11	0.91	
Mouse Tracking $\in$ {*Comfortable, Very Comfortable*}	0.76	2.13	0.54	1.40	0.16	
Keyboard Restr. $\in$ {*Comfortable, Very Comfortable*}	−0.19	0.83	0.48	−0.39	0.70	
Screen Recording $\in$ {*Comfortable, Very Comfortable*}	1.00	2.72	0.53	1.87	0.06	.
Mic Recording $\in$ {*Comfortable, Very Comfortable*}	1.11	3.04	0.67	1.66	0.10	.
Webcam Recording $\in$ {*Comfortable, Very Comfortable*}	1.96	7.08	0.63	3.11	<0.001	**
Intercepts						
Very uncomfortable | *Uncomfortable*	−1.42	0.24	0.39	−3.61	<0.001	***
Uncomfortable | *Neither comfortable nor uncomfortable*	0.59	1.81	0.33	1.81	0.07	.
Neither comfortable nor uncomfortable | *Comfortable*	1.67	5.32	0.37	4.52	<0.001	***
Comfortable | *Very comfortable*	4.17	64.80	0.61	6.89	<0.001	***

Signif. codes: '***' $\widehat{=} < 0.001$; '**' $\widehat{=} < 0.01$;'*' $\widehat{=} < 0.05$; '.' $\widehat{=} < 0.1$

Table 6: Ordinal regression model to describe the preference for online proctored exams based on responses to Question **Q12**. The model uses an ascending agreement scale (i. e., from *Strongly disagree* to *Strongly agree*). The Aldrich-Nelson pseudo R^2 of the model is 0.75.

Factor	**Estimate**	**Odds ratio**	**Error**	**t value**	**Pr(>|z|)**	
Exams taken > 3	−0.07	0.93	0.40	−0.18	0.86	
Aware methods $\in$ {*Moderately aware, Extremely aware*}	0.83	2.28	0.67	1.23	0.22	
Concern amount $\in$ {*Disagree, Strongly disagree*}	−1.76	0.17	0.81	−2.17	0.03	*
Privacy invasive $\in$ {*Disagree, Strongly disagree*}	2.21	9.10	0.66	3.35	<0.001	***
Reasonable tradeoff $\in$ {*Agree, Strongly agree*}	0.94	2.57	0.56	1.69	0.09	.
Good solution $\in$ {*Agree, Strongly agree*}	1.30	3.66	0.52	2.48	0.01	*
Comfort methods $\in$ {*Uncomfortable, Very uncomfortable*}	−0.95	0.39	0.49	−1.93	0.05	.
Concern sharing $\in$ {*Disagree, Strongly disagree*}	0.58	1.78	0.81	0.71	0.48	
Intercepts						
Strongly disagree | *Disagree*	−0.67	0.51	0.50	−1.34	0.18	
Disagree | *Neither agree nor disagree*	1.40	4.07	0.54	2.61	0.01	**
Neither agree nor disagree | *Agree*	2.39	10.90	0.57	4.17	<0.001	***
Agree | *Strongly agree*	$1.14E+02$	0.72	6.61	$3.74E-11$	***	

Signif. codes: '***' $\widehat{=} < 0.001$; '**' $\widehat{=} < 0.01$;'*' $\widehat{=} < 0.05$; '.' $\widehat{=} < 0.1$

Table 7: Post-Hoc Analysis of comfort with monitoring method **Q28** using pair-wise Mann-Whitney U-Test with Holm-Sidek Correction.
(Kruskal Wallace: $H = 94.6, p < 0.001$)

	Lockdown browser	Keyboard Restr.	Live proctor	Mouse Tracking	Screen Rec.	Webcam Rec.	Mic. Rec.	Browser Hist.
Lockdown browser	—							
Keyboard Restr.	0.553	—						
Live proctor	0.232	0.930	—					
Mouse Tracking	0.021*	0.822	0.930	—				
Screen Rec.	0.001*	0.261	0.666	0.857	—			
Webcam Rec.	< 0.001*	0.005*	0.039*	0.160	0.822	—		
Mic. Rec.	< 0.001*	0.003*	0.027*	0.095	0.764	0.930	—	
Browser Hist.	< 0.001*	< 0.001*	0.001*	0.007*	0.267	0.920	0.930	—
Eye Tracking	< 0.001*	< 0.001*	< 0.001*	< 0.001*	0.032*	0.635	0.764	0.930

Table 8: Post-Hoc Analysis of neccessity of monitoring method **Q27** using pair-wise Mann-Whitney U-Test with Holm-Sidek Correction.
(Kruskal Wallace: $H = 92.8, p < 0.001$)

	Lockdown Browser	Webcam Rec.	Screen Rec.	Live Proctor	Mic. Rec.	Browser Hist.	Keyboard Restr.	Eye Tracking
Lockdown Browser	—							
Webcam Rec.	0.605	—						
Screen Rec.	0.171	0.946	—					
Live Proctor	0.004*	0.582	0.911	—				
Mic. Rec.	0.004*	0.490	0.865	0.946	—			
Browser Hist.	< 0.001*	0.010*	0.084	0.472	0.677	—		
Keyboard Restr.	< 0.001*	< 0.001*	0.009*	0.092	0.336	0.946	—	
Eye Tracking	< 0.001*	< 0.001*	0.004*	0.044*	0.181	0.946	0.946	—
Mouse Tracking	< 0.001*	< 0.001*	< 0.001*	< 0.001*	0.010*	0.605	0.865	0.946

Virtual Classrooms and Real Harms: Remote Learning at U.S. Universities

Shaanan Cohney
Princeton University / University of Melbourne

Ross Teixeira
Princeton University

Anne Kohlbrenner
Princeton University

Arvind Narayanan
Princeton University

Mihir Kshirsagar
Princeton University

Yan Shvartzshnaider
Princeton University / York University

Madelyn Sanfilippo
Princeton University / UIUC

Abstract

Universities have been forced to rely on remote educational technology to facilitate the rapid shift to online learning. In doing so, they acquire new risks of security vulnerabilities and privacy violations. To help universities navigate this landscape, we develop a model that describes the actors, incentives, and risks, informed by surveying 49 instructors and 14 administrators at U.S. universities. Next, we develop a methodology for administrators to assess security and privacy risks of these products. We then conduct a privacy and security analysis of 23 popular platforms using a combination of sociological analyses of privacy policies and 129 state laws, alongside a technical assessment of platform software. Based on our findings, we develop recommendations for universities to mitigate the risks to their stakeholders.

1 Introduction

The COVID-19 pandemic pushed universities to adopt remote educational platforms. But most of these platforms were not designed with universities in mind. While these platforms allowed institutions to fill an urgent need, they caused novel and well-publicized security and privacy problems. We examine the underlying causes of these problems through an interdisciplinary lens that identifies the institutional structures that make these incidents more likely, and surfaces the tensions between educational goals and the incentives of the software platforms.

We begin in Section 2 by documenting how different actors in educational settings—students, instructors, and administrators—each bring their own set of preferences and concerns about platforms. These concerns conflict, leaving instructors and students frustrated that platforms do not meet their needs. We use qualitative surveys of 49 instructors and 14 administrators from U.S. universities to help model the considerations. Next, in Section 3 we discuss the risks that emerge from complex social interactions between the actors in our model through the lens of Contextual Integrity (CI) [45], and discuss where the pandemic has disrupted norms for appropriate information flows. In Section 4 we discuss the security threats that compromise digital systems through unauthorized access. We discuss related work in Section 5 and future work in Section 6. We synthesize our analyses into recommendations for universities and regulators in Section 7.

Our analysis identifies three factors that contribute to privacy and security problems. First, there are unresolved tensions between the needs of the different stakeholders. For example, instructors' stated preferences for students' cameras to be left on conflicts with students' privacy concerns about misuse of their video feeds. Second, there are significant gaps between users' preferences for platform behavior and the actual practices of the platform. A recurring theme in our survey was that defaults matter. While developers may include a configuration toggle for stricter security settings (such as storing call recordings locally vs. to a cloud), if this toggle is not turned on by default or communicated clearly to instructors, its usefulness is significantly hampered. A related issue is that faculty often adopt tools on their own initiative, outside official procurement processes which deprives them of protections afforded to large organizations. Third, there is a regulatory gap created because the existing educational privacy and data security regulations were written for an era of paper-based records in physical classrooms and are not a good fit for regulating practices arising out of remote learning.

Contributions:

- We build a novel threat model that represents the stakeholders in virtual classrooms, their interactions, and pri-

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

vacy/security risks. We ground our model in two qualitative surveys we conduct of instructors and college and IT administrators at U.S. universities.

- We assess the privacy and security practices of 23 of the most popular platforms identified in our survey. In particular, we find notable differences between the practices of platforms operated under contract with universities, and the same platforms provided to users free of charge. We observe that contracts negotiated with universities result in significant differences in how data is handled by the platforms. We find that these Data Protection Addenda (DPAs) are a powerful tool to shape platform behavior towards the interests of the stakeholders.

- We use our research to provide policy guidance. We recommend that universities prioritize developing tools to incorporate continual improvements based on user feedback and allowing instructors to select features that are relevant to individual educational missions and protecting the interests of vulnerable groups. We also recommend strengthening regulatory mechanisms to provide appropriate baseline privacy and security protections.

2 Actors, Incentives, & Risks

In education technology, the actors within the system are both principals to protect and potential threats to mitigate. A useful threat model should therefore specify the level of trust to assign to each actor, in recognition of the variety of roles that individuals in that class may play.

Our construction of a threat model is further complicated by the fact that a platform component may seem both a 'feature' and a 'threat' to different actors (e.g. video recording). Platform developers must not only build a secure product, but one that mediates between the different interests of their users. As a result, our model moves beyond a trusted/untrusted dichotomy to examine the incentives and interests of all actors.

We begin by describing our threat model to understand how the participating actors view their interests. Next, we model the different risks that platforms must mitigate to help fulfill the educational mission of the software. Finally, we conclude by analyzing the survey results that inform our threat model. Our threat model is based on frameworks such as STRIDE [38], socio-technical system analysis [41], and Contextual Integrity [45].

2.1 Actors

We divide the participating actors into internal actors (students, instructors, administrators), and external actors (third parties and adversaries). While each *class* of internal actors has incentive to maintain the integrity of the educational system, we recognize that all internal actors may engage in adversarial behavior. In addition, our survey shows how the priorities of the internal actors may differ—leading to behavior by one internal actor which may be considered adversarial by another actor.

Incentives. Students enroll for reasons beyond the pursuit of education goals. Further, diplomas have a credentialing function which, for some students, may incentivize cheating. Mixed incentives may therefore cause a student to act contrary to other stakeholders' interests.

Administrators who do not personally use the digital classroom may be more willing to sacrifice the privacy concerns of students and instructors for institutional concerns such as cost efficiencies and auditing or reporting capabilities. Administrators commonly endorse cloud solutions such as Canvas which monetize aggregate data about students, representative of trade-offs between different actors' priorities.

While our model groups instructors and administrators together, they are not homogeneous. Instructors and administrators may play different roles in configuring and using platforms. Those administrators whose roles focus on compliance and risk assessment have incentives to prioritize security and privacy concerns, while others prioritize usability and teaching outcomes. At different institutions, various groups may participate in platform procurement, including institution-level administrators, department staff, and instructors, and power dynamics between these groups may also differ. This also highlights a limitation of our model, which draws boundaries between instructors, administrators, and students, when there are often instances in which those roles overlap—for example, graduate students who teach or TA or senior faculty who serve as administrators.

Under these assumptions, we produce the following set of descriptions about the behaviors of our actors:

- **Students** authenticate their identities and participate in courses by joining live virtual sessions and submitting work. Students may also collaborate and share work with each other. Students may share extraneous information—including their home environment—through video conferences that propagate to instructors (and which may leak to platforms if conferences are not end-to-end encrypted). We do not assume that students are trusted: actors with student-level permissions may try to access or change data not authorized for them, or to prevent access to systems. We also note that students may participate in privacy violations [40]. Violations may be intentional when students exploit data-rich platforms or unintentional when platforms leak sensitive information (such as indicators of socio-economic status leaked via video streams).

- **Instructors and administrative staff** generally manage instances of the various platforms (such as individual chat rooms/streams), and are thus significantly privileged and trusted. Well-intentioned instructors may inadvertently breach student or institutional expectations of

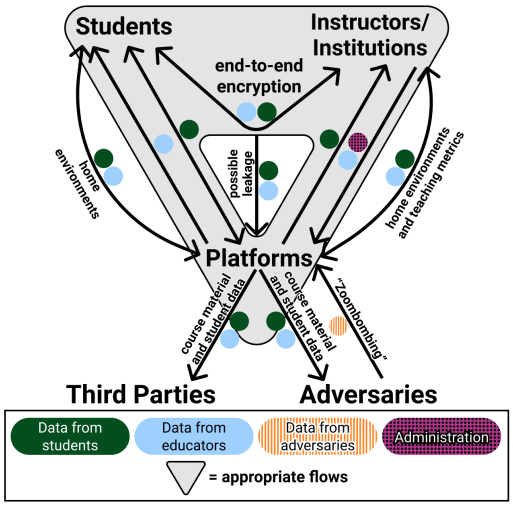


Figure 1: We model typical data flows between stakeholders as mediated by online platforms. A flow is considered appropriate if it corresponds to a legitimate flow under the Contextual Integrity framework we develop in Section 3. Data from end-to-end encrypted sessions may leak to platform due to poor configuration, implementation, or metadata collection.

privacy in a virtual classroom, while misinformed administrators may improperly use student data or metadata to harm students [21], such as through false accusations of cheating [64]. Similarly to students, instructors can precipitate security breaches if their accounts are over privileged, and may prefer to keep extraneous information from being shared with students and administrators (including their own home environments and teaching metrics).

- **Service providers and their third-party affiliates** act to maximize their economic interest within the bounds of their contractual and legal obligations. Platforms may share metadata with third parties for advertising and other business purposes, and course material for services like captioning.

- **External adversaries** may seek to steal student data and course content, and may interfere with live classes ("Zoombombing"). Adversaries may act for profit, entertainment, or other motives.

Figure 1 depicts these actors and their interactions.

2.2 Survey

We built our threat model using a survey of 49 instructors at U.S. universities to learn what remote learning platforms they use, the features they value in a platform, and their concerns (and those of their students), with particular emphasis on privacy and security. We recruited participants through a public Slack group for instructors teaching remotely, as well a public social media post.

Separately, we surveyed 14 U.S. university administrators about their schools' procurement processes for new learning platforms. As administrators with influence and understanding of the procurement process are harder to reach, our sample size was limited to 14 participants.

Our institutional review board (IRB) reviewed and approved our study design, consent, and recruitment procedures. All participants affirmatively consented to participate in the study, after reviewing a form approved by the IRB.

The surveys provide a qualitative framing for our platform analyses and recommendations later in the paper. Additional details regarding survey materials and responses are provided in Appendix A.

2.2.1 Instructor Survey Results

We report the number of instructors using each platform, as well as whether they use a personal version or an institutionally provided version in Figure 2.

We find that the concerns of instructors generally touch on three major themes, with some instructors mentioning multiple themes. First, students' personal data is more easily captured and visible to instructors, platforms, co-inhabitants, and other students from a virtual classroom. This includes students broadcasting their home environments (and socioeconomic indicators therein) through video, private chats which may leak to instructors, and co-inhabitants hearing sensitive class material. It also includes proctoring services that hijack students' computers while monitoring their environment.

Second, personal data is more easily disseminated by platforms to third parties, with little recourse for students or instructors who wish to limit data sharing. Instructors were concerned that platforms may sell metadata from students/instructors to advertisers or leak data to services like video captioning, and that law enforcement may request student data from platforms.

Third, platforms are vulnerable to attack from unintended adversaries that threaten to steal data and interfere with courses, due to poor authentication and other security measures by platforms.

Specific concerns and desires mentioned in our instructor survey are listed below.

Security/privacy. 25 (51%) instructors marked "Yes" to having security and/or privacy concerns with platforms. Respondents did not clearly delineate between security and privacy: notably, each of our five freeform questions, including those unrelated to privacy or security, received at least two responses that we manually coded as pertaining to a privacy or security concern. Fifteen discussed concerns that "private" chats or videos were not private, especially to other students; nine were concerned that platforms did not adequately secure meetings against intrusion; and two mentioned concerns

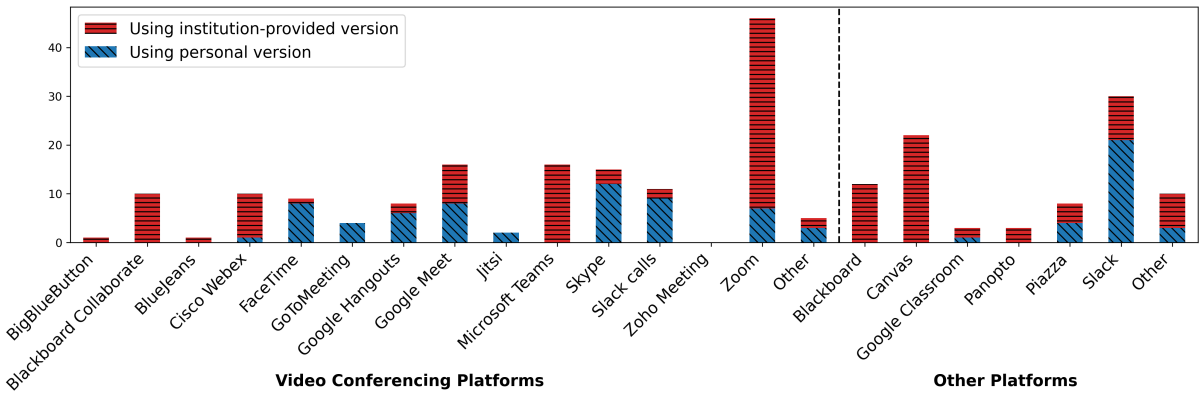


Figure 2: Survey usage counts for remote teaching platforms, including institution-supported versions and personal versions.

about theft of intellectual property.

Surveillance. Instructors were concerned about the surveillance implications of using remote learning tools, both for student data and their own. For example, five instructors noted they were concerned platforms might share data with third parties, with particular emphasis on course data and personal data of students. Another instructor also denounced potential data sharing with law enforcement. Meanwhile, two instructors mentioned not wanting to share their own home environment with students, and one was concerned that their campus would use platform metrics to judge their teaching performance. One instructor even tried to use Privacy Badger [28] to disable tracking on Blackboard, but this interfered with classroom functions.

Recording restrictions. Nine instructors reported it was important that platforms provided the ability to save recordings locally or to private clouds. One advocated for allowing students to "opt out" of showing their video on recordings. Together, these respondents discussed restrictions on every aspect of class recordings: restricting who can make recordings, who hosts them, and who is allowed to access them.

Platform choice. Respondents also reported specific dissatisfaction with their institutions in selecting and configuring platforms. Seven instructors reported frustration with the choice of platforms by their institutions, with one instructor noting that Canvas' 'testing capabilities are not as nice as' a platform used at another university. Two of these instructors discussed using alternative, free software such as Humanities Commons and Mattermost. The other instructor noted a stored message limit in the free version of Slack, which demonstrates an issue for instructors using software that is not supported by their institutions. One instructor also reported "We don't have the [Canvas] version with video conferencing," whereas conferences are a configurable permission by universities at no extra cost [18]. This highlights that instructors and universities may not be aware of configurable or default settings in platforms.

2.2.2 Administrator Survey Results

While our administrator survey had only 14 respondents, it surfaced issues about how their institutions select software—issues other universities are likely to share. Eight administrators reported that their institutions did not have a formal process for selecting new platforms, while one reported a price threshold for invoking a more formal process. Administrators marked that the platform selection process was driven primarily by Faculty and IT staff. Two administrators reported having processes for collecting feedback from faculty on learning platforms, but one emphasized that re-evaluation of whether a platform met needs would only happen at the time of "contract renewal." Further, "legal obligations" were rated moderately important (when rated from 'not at all' to 'extremely' important) by 3/4 administrators who answered this question, implying that other actors are likely responsible for compliance. Four administrators rated a platform's privacy features as 'very' or 'extremely' important. Three of those four rated security equally important, while one rated security only 'slightly' important. Three administrators described negotiating for more privacy guarantees such as DUO and HIPAA compliance.

According to one administrator, the most significant change relative to COVID-19 was not procurement of additional video platforms, but filling new needs such as proctoring software for "high-stakes testing."

3 Privacy Analysis

As the cost of data collection from online interactions is low and there is less friction to collect such data compared to the offline context, the platforms end up collecting vast amounts of data. As discussed below, these practices, despite ostensibly being compliant with the existing regulations, can conflict with with contextual educational privacy norms and expectations.

3.1 Empirical Approach to Privacy Analysis

Given that the expectations and interests of the relevant actors are complex, conflicting, and overlapping (as discussed relative to incentives in Section 2.2), we adopt descriptive institutional analysis frameworks [31,58] to structure our governance inquiries. This approach recognizes and builds upon a conceptualization of technology governance as an assemblage of laws, norms, markets, and architecture [30,35,39]. We also use the Contextual Integrity framework to understand how stakeholder expectations change when the physical classroom becomes digital, drawing on established methods [63]. CI views privacy as the appropriate flow of information, where appropriateness is defined by the governing contextual norms. We draw on the survey responses to identify which information handling practices are appropriate in the educational context.

We employ the existing integrated GKC-CI codebook [59] to operationalize these frameworks and assess governance of information flows associated with the platforms. Two of the investigators assessed inter-rater reliability in two phases, based on a subset of privacy policies. After the first round, overall agreement was 86.27% with ranges in Krippendorf's alpha from .49 to .97, with 2 of 8 codes not reaching the required .8 threshold. Following inter-rater discussion, a revised second round of coding was conducted wherein agreement improved to 93.67% overall, with Krippendorf's alpha ranging from .81 to .97, indicating excellent agreement. The same investigators subsequently applied the codebook to information flows and governance described in DPAs and regulations.

3.2 Law

We observe that current laws do not sufficiently control platform behavior to conform to the privacy norms of higher education. While the market for educational technology is comparatively highly regulated by state and federal laws, those laws are not always effective. For a university to control a platform's information practices in a way that fits within the spirit (if not the actual application) of federal and state laws, it must take active and intentional supplementary governance interventions, such as by customizing DPA as we discuss in Section 3.1.

3.2.1 Background

We provide a brief overview of the primary legal frameworks in the United States that apply specifically to student privacy.

Federal Regulation. The Family Educational Rights and Privacy Act (FERPA) [9] protects defined categories of student records and enrollment at an educational institution [52]. The law, enacted in 1974, was designed for a paper-based record system with discrete and limited set of records.

FERPA requires schools and universities to keep records of each external disclosure of student information and requires the records to be available on request by the subject [9]. FERPA regulates information sharing by requiring that the institution gets explicit affirmative consent to share data with third parties that fall outside listed exceptions. If the documentation a platform provides does not concretely describe how it shares user data with partners or advertisers, it may run afoul of the regulation.

FERPA only applies to organizations that receive federal funds under certain educational programs. It is mainly enforced when the Department of Education determines that a school or university is in violation. The department then enforces FERPA by withholding federal funds until they come back into compliance.

FERPA specifies what student information can be shared with whom, distinguishing between situations in which consent is required, and those in which it is not. The act permits schools to disclose records to contractors under certain conditions: third-parties must be under the educational institution's direct control, and must be designated as school officials having "legitimate educational interests."

FERPA's other notable allowance of data sharing is for directory information—"name, address, telephone number, date and place of birth, honors and awards, and dates of attendance"—which can be shared so long as adequate notice and opportunity to request non-disclosure is provided [9]. Note that Universities must maintain records of when they share directory information sharing. In all other instances, explicit consent is required, corresponding to a norm that privileges student privacy without informed consent. FERPA reinforces this norm throughout its provisions on disclosure and consent.

While broad in scope, FERPA is limited to a general set of expectations that addresses specific categories of *information types* and *information recipients*. But the rules and formal norms do not translate well to the digital environment, and do not specify transmission principles—when transmitting data is appropriate or who are permissible senders or recipients of data transmissions [75]. In particular, FERPA does not supply any specific guidance about what educational technology platforms can do with the data they generate and collect about students. However as "designated school officials" they may only share that data with other such officials. Other third-party sharing is not permissible, with obligations and limits specified in Department of Education Guidance, originally drafted under the Obama administration and currently applied under the Biden administration [53].

State Regulations. State privacy legislation affects platform practices. This includes both general privacy laws (most prominently the California Consumer Privacy Act (CCPA) [10]) as well as specific laws that regulate student privacy. Specifically, 45 states (which for our purposes includes Washington D.C.) have more than 129 educational privacy laws [13,23], some of which regulate school and student interaction or data collection by digital platforms [13].

Many state laws take inspiration from California's Student Online Personal Information Protection Act (SOPIPA) [5] of 2014, which was intended to comprehensively cover K-12 student privacy concerns. In contrast to FERPA, SOPIPA imposed liability on platforms and providers, in addition to schools.

We aggregated 129 state educational privacy laws, as tracked by Student Privacy Compass [23] and the Center for Democracy and Technology (CDT) [13] and coded them to identify and compare information flows and governance patterns, through a combination of manual and hybrid tagging, drawing on established methodologies [26, 62]. We present summary data and publish the corpus alongside this work at https://github.com/edtech-corpus/corpus.

Almost all states in the corpus had laws that required significant transparency about data sharing practices. 5 states allowed students and families to opt-out of personal information sharing across the board without making a case-by-case determination. 11 states require affirmative consent, opting-in, to share some categories of or all personal information with any recipients outside the school district. 21 state laws included bans on targeted advertising. 6 states were not present in our data set as they had not passed any applicable student privacy laws.

3.2.2 Analysis

Regulation at the state level is often more precise than FERPA in addressing specific aspects of digital information flows, limiting platforms as information senders and recipients, as well as articulating clearer transmission principles. Although state level regulations tend to specify more details with respect to permitted information flows in and out of the respective platforms, this layer of governance varies across states and places the burden of compliance on universities rather than the platforms. Moreover, these state laws were primarily designed for the the K-12 context, leaving substantial gaps in the regulatory framework. This mode of privacy regulation imposes more relevant institutions, but is still limited in pertaining to a subset of relevant information subjects and varies significantly from place to place, with schools and universities bearing the burden of compliance, rather than providing a common floor for minimum protection.

A more pervasive issue underlying state or federal laws is that they have limited enforcement mechanisms or penalties for misconduct. For example, under state laws there are no "private rights of action", meaning that the laws did not grant students or their guardians the right to sue if a provider violates the law. Primary enforcement is left to state attorneys general, who have limited resources to pursue breaches. Thus, there are few incentives to police compliance with state legal requirements.

As FERPA does not regulate how platforms use data, focusing instead on schools and universities, platforms used in higher education have leeway to use and abuse educational data once it enters their custody.

Universities can fill gaps by introducing their own polices and rules, as well as by extracting binding commitments from commercial partners through contracts. We explore use of these binding commitments in Section 3.4.

- Apple Classroom
  Apple Facetime
  Apple Schoolwork
- BigBlueButton
- Blackboard
- Blackboard Collaborate
- BlueJeans
- Canvas
- Jitsi
- G Suite for Education
  Google Classroom
- Google Hangouts
  Google Meet
- GoToMeeting
- Microsoft Teams
  Microsoft Skype
- Microsoft Skype for Business
- Panopto
- Piazza
- Slack
- WebEx Meetings
- Zoho Meeting
- Zoom

Table 1: The 23 platforms whose policies we examined. There were fewer policies than products, as some firms (such as Microsoft) have monolithic policies that apply to groups of products.

3.3 Privacy Policies

Platforms self-regulate through self-imposed privacy policies, in which they structure and disclose sharing with third parties. Privacy policies may restrict information flows while containing broad language that hedges on specifics. Common sources of flows to third parties include integration between platforms or sharing of data for analytics and marketing.

We manually coded privacy policies for 23 platforms, which corresponded to 18 integrated policies as shown in Table 1.

Of the platforms, 13 were mainly available as enterprise products, typically requiring institutional support, and 10 could be adopted at will by individual instructors.

We followed the methodology in [62] to annotate statements in each policy based on CI parameters to classify and describe information flows between users, platforms, and third-party entities. For example, in the following quote from Zoho's privacy policy:

> "We collect information about you only if we need the information for some legitimate purpose."

the pronouns "We" and "you" are labeled as Receiver (the service provider) and Subject (user) of "information", respectively. We label "for some legitimate purpose" as transmission principle, i.e., the condition under which the information is

Description	Frequency
Third Party Sharing	
Burden on users to monitor third-parties	8 (44%)
May share personal data with advertisers	8 (44%)
Bi-directional sharing	6 (33%)
May collect personal data from social media	7 (38%)
Location Sharing	
Explicitly permit location tracking	10 (55%)
May share location data with third-parties	4 (22%)
Collect location data outside device-provided	5 (27%)

Table 2: We identified the privacy practices of 23 platforms from 18 different privacy policies. There are fewer policies than platforms as products owned by a common firm typically shared a policy, indicated by a single bullet spanning multiple platforms.

being transferred. Note that the statement does not specify the sender of the information.

The policies we collected serve both as a source of empirical information about patterns in platform practices and a series of case studies that reflect differing governance practices. We summarize our results in Table 2, and present expanded results in Appendix B.

Third Party Sharing. Eight of 18 platform policies explicitly informed users that the burden was on the user to monitor third party firms whose products were integrated with the primary platform. Three platforms specified that agreements with third-party providers provided some privacy protections. The remaining 11 were unclear about third party sharing.

Where a policy applied to EU citizens, the text would typically specify that third parties were also bound to the protections offered by the platform.

While sharing an ID may seem innocuous, student IDs have long served multiple functions, many of them security sensitive. Blackboard's integration policy permitted Blackboard to share school-provided student IDs with partners.

Eight of 18 platform policies allow the platform to share personal information with advertisers and marketers. Three of 18 policies contained inconclusive language. Only 2 platforms did not share personal data with third parties for any purpose other than those mandated by law.

Six policies allowed bi-directional sharing. For example, BlueJeans' policy permits collection of user data *from* "other Service users, third-party service providers...resellers, distributors, your employer, your administrator, publicly available sources, data enrichment vendors, payment and delivery service vendors, advertising networks, analytics providers, and our business partners"—a list that incorporates any conceivable third party.

Seven policies allowed platforms to collect user information from social media, with Zoho going even noting that "once collected, this information may remain with us even if you delete it from the social media sites."

Slack's policy, like those of many platforms, places the burden on users to "check the permissions, privacy settings, and notices for... third-party Services" whom Slack may receive data from, and to "contact [Services] for any questions."

We found significant variation in the level of detail among privacy policies, with only a minority of policies offering detailing specifically when, to whom, and under what conditions information is shared.

Location Sharing. Of the 18 policies we evaluated, 12 policies permitted location tracking, 5 explicitly stated they did not track location, and 1 was unclear.

Of the 12 that collected location data, 4 policies allowed data sharing with third parties. Reasons for sharing and uses permitted varied from the relatively benign (sharing to a mapping company for displaying maps) to the worrisome (sharing for marketing and advertising). Other policies provided broad discretion for uses of anonymized location data. Among the policies with broad language were Google and Apple, whose policy allowed them to share location data with "partners and licensees to provide and improve location-based products and services."

Six of 18 policies mentioned capturing location data using mechanisms other than mobile-device provided location. Notable examples were Slack, which approximated location using information gathered from third parties, and Google, which referenced search data. Four policies did not disclose how they implemented location tracking.

Many policies did not clearly explain why they collected location data, beyond minimal examples under the umbrella category "improving our services." Moreover, the language of the privacy policies could encompass uses that instructors and students might object to—such as Piazza's policy which permits uses "as required or permitted by law."

3.3.1 Analysis

Our results show that the the governance of platforms and the needs of our stakeholders are not aligned by default. For that reason, considerable negotiation or governance is necessary at the university level to platforms' behavior with our stakeholders' needs.

Our finding that some platforms use a broader range of tracking techniques beyond device-provided location services strips choice away from users. By bypassing device-based restrictions on obtaining location data, platforms subvert users' expectations.

As advertising and marketing third parties are integral parts of the digital economy it is unsurprising that many apps we examined interact with third-parties for marketing or advertising purposes. Policies often failed to enumerate the categories that constituted personal information, leaving platforms with broad discretion to what is appropriate to share.

One notable finding was that platform sharing with third

parties was in some instances bidirectional—platforms received user data from social networks and other parties, while at the same time transmitting user data to these parties. Platforms may thus be able to build profiles of their users in ways that violate student and institutional expectations, which are derived without this knowledge.

Privacy policies generally reflect defaults applied to individually licensed versions of these tools (which are free or low-cost), reflecting norms in the sense of Lessig's model for governance [35, 39]. Institutions can negotiate provisions when they engage contractually with platforms. But, when individual instructors use these platforms, they do not always realize that free or default licenses do not meet regulatory or normative expectations for privacy protections.

3.4 Data Protection Addenda

By leveraging their status as large organizations, universities can negotiate commitments with platforms, called Data Protection Addenda (DPAs), that specify local rules and characterize additional responsibilities and expectations for institutionally supported platforms.

The DPAs we analyzed reflect three distinct types of contractual relationships: one-to-one, one platform to many universities, and one university to many platforms. Platforms offer template DPAs to make it easier for enterprises, including hospitals and universities, to adopt a given platform. Zoom provides their own templates, emphasizing FERPA and HIPAA obligations, as do Microsoft Teams, Google Hangouts, and Skype for Business. Other types of DPAs include those negotiated between specific universities and specific platforms and those drafted by individual universities and applicable to all vendors. The differences between public universities that negotiate their own agreements and those that use templates are not obviously correlated with factors such as endowment or student body size.

We coded 50 publicly-available DPAs from a cross section of 41 public universities and 4 private universities. Many universities in our dataset also appear to have other non-public agreements, including with some of the same platforms, as described in the DPAs analyzed and in public FAQs.

Eleven of the 50 agreements negotiated different sets of rules for educational, organizational, human subject research, and medical uses (relative to university hospital use) within the same document. 10 of these 11 specifically differentiate between educational or enterprise media data and additional protections or scrutiny for university hospital data, such as the documents negotiated between Zoom and the University of Florida, or WebEx and Iowa State University. Another notable modification was Zoom's commitment to allow the University of Illinois to self-host the platform.

4 DPAs for Zoom and 2 DPAs for WebEx were consistent across 6 different universities, including the University of Minnesota and the University of Pittsburgh, implying the platforms' suggested DPAs were employed. In contrast, 18 universities had unique DPAs that correspond with multiple platforms and vendors. We also found significant variation in access to data and duration of data retention. For example, under the DPA between Zoom and the University of Virginia, Zoom's obligations "survive termination...until all University Data has been returned or Securely Destroyed".

The DPAs negotiated by the University of California exhibit similarity as they are designed to comply with the University's Electronic Communications Policy [68], showing the impact of local policy. The same constraints can be seen in the DPAs negotiated by Florida State University, which employs information classification guidelines drawn from Florida's public record laws. As a result FSU's agreements include consistent language and requirements for all platforms through which student data is collected, stored, or processed [6]. This shows how state regulations can shape behavior, even when they don't apply directly.

For the University of Connecticut (UConn), a DPA negotiated by the state with a platform applies not only to the university but to all other public institutions. Besides restraining platforms' practices, rules can impact how universities select features and defaults. In the case of UConn and Zoom, the agreement also places expectations on users (such as obligations to use a passcode and regularly update software), which are enforced by platform settings [24].

Software	Version
WINDOWS/MACOS	
Zoom	4.6.10
Slack	4.5.0 (64bit) / 4.4.2
BlueJeans	2.19.791 / 2.19.2.128
Jitsi	2.10.5550
Cisco WebEx Meetings	40.2.16.14
Cisco WebEx Teams	1.0.0.2
Microsoft Teams	1.3.0.8663

Table 3: **Software Evaluated in this study (Desktop Applications)** We evaluate a set of commonly used platforms in remote learning environments. Separate Windows/macOS version numbers are given, where necessary, throughout this work.

4 Security Analysis

We address the paucity of existing security analyses by performing a deeper analysis of the top video-conferencing and collaboration tools, limiting this portion of our analysis to desktop software (shown in Table 3). We additionally provide a short analysis of mobile app permissions in Appendix C.

Our analysis spans four metrics, chosen to maximize ease for an administrator to replicate our procedures.

	Zoom	Slack	BlueJeans	Jitsi	WebEx (M)	WebEx (T)	MS Teams
				WINDOWS/MACOS			
Arch	i386/AMD64	AMD64	AMD64	AMD64	i386/AMD64	AMD64 / AMD64	AMD64
SafeSEH	✗	N/A	N/A	N/A	✓	N/A	N/A
DEP/NX	✓/✓	✓/✓	✓/✓	✗/✓	✓/✓	✓/✓	✓/✓
ASLR	Low / ✓	High / ✓	High / ✓	✗	Low / ✓	High / ✓	High / ✓
CFI	✗	✓	✗	✗	✗	✗	✓
Code Signing	✓/✓	✓/✓	✓/✓	✓/✓	✓/✓	✓/✓	✓/✓
Stack Canaries	✓/✓	✓/✓	✓/✓	✗/✓	✗/✓	✗/✓	✗/✓

Table 4: Security features present in end-user binary software on Windows and Mac desktop OSes. Where a feature is available for both Windows and macOS, support in the for the feature is marked for Windows/Mac on the left and right of the slash respectively. Low/High ASLR entropy indicates whether compile time flags necessary for high entropy ASLR were present. Rows with only a single element per entry represent features specific to Windows. As SafeSEH applied only to 32-bit binaries, 64-bit binaries have their respective entries marked N/A. WebEx Teams/Meetings are distinguished by (T) and (M) respectively.

4.1 Network Traffic Analysis

We captured network traffic to and from desktop software packages from each platform to determine with whom the application communicated and whether and how it encrypted these flows. We used application-layer traffic analysis in two ways: to search for any qualitative security red-flags (such as obvious failures to encrypt data-in-transit, poor choice of TLS cipher suites), and to look for flows to third parties.

We performed all captures on a clean Windows install using the same software packages profiled in our binary analysis (Section 4.2). We used a monster-in-the-middle attack to interpose between the software packages and the servers they were contacting, allowing us to see unencrypted traffic.

We began each capture, then started and logged in to the software being tested, began a video call with a second client (for all applicable platforms), terminated the call, then terminated the capture. We counted the number of third-party domains to which each connected and looked for domains with no apparent connection to the provision of platforms' services. We also ran the Qualys SSL Labs tests [54] against servers to which client software connected. We excluded all domains that we could identify with the platform operator (e.g.; `slack-edge.com`).

We included third-parties that may add platform functionality, such as Gravatar (graphic avatars) or Amazon Chime (video conferencing). Though use of these services may often be justifiable, including them in our results contributes to an understanding of how broadly platforms may share user data.

Results

Only BlueJeans and Slack connected to third party hostnames. BlueJeans connected to New Relic, Microsoft and MixPanel, all for analytics. Slack connected to Gravatar and Amazon Chime, which Slack uses for video calls. The full set of domains are given in Appendix D.

Slack, WebEx Teams, and Microsoft Teams all used certificate pinning to verify the identity of the servers they connect to, providing protection against TLS monster-in-the-middle attacks. WebEx Meetings and Zoom presented warnings for untrusted certificates, but allowed users to click through the warning dialogs. If a user clicks through the Zoom warning, Zoom will persistently trust the certificate across executions.

Client software generally requested safe TLS cipher suites, as classified by SSL Labs, but unfortunately all the platforms maintained support for RSA suites, which are known to be weak and vulnerable to attack. Bluejeans and Jitsi supported finite-field Diffie-Hellman cipher suites that researchers similarly warn against [69]. The cipher suites offered by platforms' corresponding servers deviated from from those of their clients, which is intriguing as—had significant thought gone into the choice of suites—the providers would have been able to ensure close matches.

We profiled servers against SSL Labs and found that all platforms' servers received scores of A or higher for all platforms except for Bluejeans, which received a B for weak cipher suites. Jitsi requires users to host their own servers and does not provide any, and so was excluded from this analysis.

4.2 Binary Security

We evaluate desktop software packages of platforms by building on the *Safety Feature* evaluation criteria of Cyber Independent Testing Labs (Cyber ITL) [27], a nonprofit research organization that attempts to provide consumer friendly security analysis of software and devices. Their approach aims to measure the difficulty "for an attacker to find a new exploit" in a given piece of software. None of these *features* impose substantial performance penalties and their absence is therefore better explained by ignorance or lack of investment in security.

The full descriptions of the features we analyze are provided in Appendix E.

Results

We extracted relevant fields from the first loaded binary image from each software package using tools provided by Microsoft/Apple where available. Where unavailable or when searching for Stack Canaries, we reverse-engineered the binaries by hand. We present our results in Table 4.

Limitations. Although the results for Jitsi appear below-par compared to the other applications, they are partially an artifact of our methodology. Jitsi is mainly written in Java, except for the main binary which uses native code to initialize the Java Virtual Machine (JVM). Our methodology reveals only the properties of this launcher and not of the underlying JVM present on the users' system. The Oracle JVM (the predominant instantiation) has well-studied security properties and protections beyond the launcher's. We therefore do not consider Jitsi's results to suggest overall poor security.

4.3 Known Vulnerabilities and Bug Bounties

Reports of software failure or flaws in the past predict failure in the future [17, 33]. We therefore analyze publicly disclosed platform vulnerabilities. We also collate and discuss the platforms' public vulnerability disclosure programs (VDPs), which are an important mechanism to aid firms in detecting and remediating software security flaws [70].

Vulnerability Disclosure Programs. Often known as 'bug bounty' programs, VDPs provide a mechanism for participants to submit flaws to a platform security team, which then (often) fixes the flaw. The programs generally offer rewards for participants—fame, fortune, or both.

Zoom and Slack both outsource their VDPs to HackerOne [2, 3], a for-profit operator that has faced criticism for its use of non-disclosure agreements that limit when a reporter may disclose the existence of vulnerabilities [50]. Zoom excluded from its program many types of potential issues including 'Attacks requiring MITM' and 'Any activity that could lead to the disruption of our service (DoS)'. Following criticism, Zoom hired external consultants to revamp its VDP [19], a process that concluded in July 2020. While Slack has a similar list of exclusions to Zoom, Slack marked them as '*unlikely* to be eligible,' leaving room for discretion.

BlueJeans outsources its VDP to Bugcrowd [1]. The only significant limitations to its rules-of-engagement are denial of service attacks, and attacks on physical infrastructure or persons. Bluejeans also provides a mechanism for testers to obtain enterprise accounts.

Finally, Cisco and Microsoft retain in-house Product Security Incident Response Teams (PSIRTs) that handle disclosure. Cisco explicitly includes high-impact vulnerabilities in *third-party* libraries used by their products in their VDP [20].

Known Vulnerabilities. When a vulnerability in software is publicly identified, it is often assigned a number according to the Common Vulnerabilities and Exposures (CVE) system.

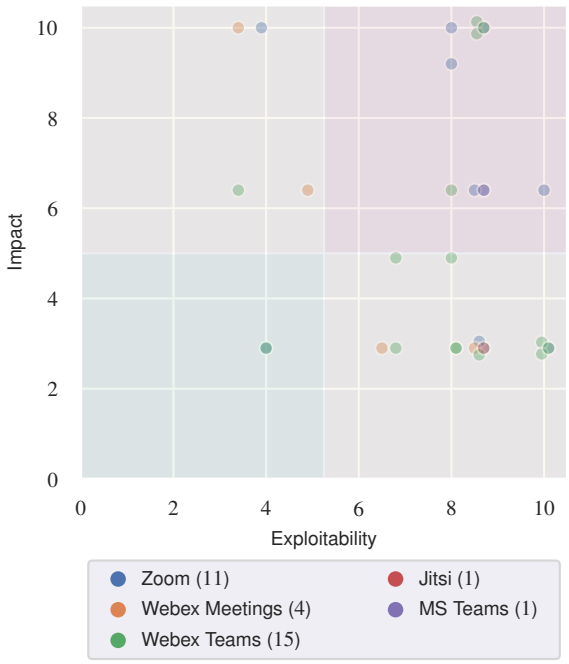


Figure 3: Exploitability and Impact scores for CVEs reported Jan 2010-May 2020. Circular clusters indicate CVEs with the same scores, with markers adjusted for visibility. Regions are shaded to indicate low/high exploitability and impact. Numbers in parentheses indicate the number of CVEs found for a particular software. No CVEs were found for Slack or Bluejeans. ($n = 32$)

A CVE ID is stored along with details of the vulnerability in the National Vulnerability Database (NVD), a separate but related program administered by the National Institute of Standards and Technology. The NVD listing includes numerical scores from zero to ten for the impact of the vulnerability when exploited, and for the ease with which the vulnerability can be exploited. These scores as calculated according to the Common Vulnerability Scoring System (CVSS).

In Figure 3 we depict the CVSSv2 impact and exploitability scores for CVEs tied to the software we evaluated. While we intended to aggregate all CVEs from 2010 onward, the earliest CVE we found for our dataset was issued in 2014—indicative of the relative newness of the platforms evaluated. The vulnerabilities are clustered toward the high-exploitability region of the figure, however this is unsurprising given that the mean impact and exploitability scores across all CVEs reported since 2010 were 8.04 and 5.04 respectively.

Zoom has many recent CVEs (11). While intense recent attention is no doubt a contributing factor, the substantial number of recent vulnerabilities suggests a systemic component to Zoom's security issues. Further, as our evaluation postdated Zoom's efforts to remediate the aforementioned security issues our results likely understate recent problems with the

software. On the other hand, Zoom's rapid improvement in both software and process (that represented a response to disfavorable media coverage) point to a positive trajectory for Zoom.

While Zoom garnered the bulk of negative media attention, in the 2019 alone, five CVEs were issued for WebEx Teams, two of which were high severity each scoring 8.6 on exploitability and 10 on impact. WebEx was issued 15 CVEs total in the reporting period, the most of all software.

Limitations. As many bugs are discovered internally, it is common for vulnerabilities *not* to be assigned a CVE ID, limiting the extent to which the number of CVEs for software can be used as a proxy for security. For example, during this study Jitsi (which only had one CVE) issued a security bulletin identifying high-severity remote execution bugs in the client software [7]. Similarly, in August 2020 HackerOne published a *critical severity* bug in Slack [47]. As of writing, these bugs were not assigned CVEs, despite their significance.

The relative number of CVEs found by researchers may also be more reflective of the scrutiny that has been applied to the platform, more-so than the quality of its software as compared to its peers.

Further, CVEs allocated to software do not reflect security issues with third-party components that developers may include and interface with. A more complete audit of would analyze all included components and vulnerabilities that may be present therein. Exemplifying this concern, in April 2020, Schroder [60] found that Zoom was using outdated libraries that contained known vulnerabilities. Flaws of this type are not reflected in our data.

5 Related Work

COVID-19 has led to a flurry of small scale remote learning and working software privacy evaluations in non-academic contexts, with notable efforts from advocacy organizations [16, 46]. Zoom has been a particular focus of many such evaluations [12, 15, 43, 60]. Concurrent to this work, the National Security Agency (NSA) published a guide to and assessment of remote collaboration software for U.S. government employees evaluating many of the same tools [8]. Their report relied on product specifications and limited technical observations. Of all products, Zoom has come under particular scrutiny from security researchers who discovered that, contrary to its marketing materials [4], it used insecure cipher modes [43], did not support end-to-end encryption, and routed users to key-servers in China [15, 43]. In response to negative media attention Zoom appears to have remediated these issues.

Prior to the COVID-19 pandemic, several academic efforts have examined the security and privacy implications of technologies deployed in education context, as well as the complexity of educational technology procurement [44]. In [72], Weller provides a historical narrative charting developments in ed-tech, noting that much of the development is due to instructors adopting broader technologies, rather than purpose-built innovation. Balash *et al.* provide the only such evaluation conducted during the pandemic, with a focus on remote proctoring [14].

The uptake of ed-tech in schools before the pandemic was on the rise. Gray *et al.* [34] performed a comprehensive survey of ed-tech in U.S public schools, finding that as of 2010 between 13% of teachers regularly used video conferencing technology in the classroom (Tab. 3) and 44% used software based testing tools (Tab. 6). They also found that 39% of students accessed a teacher's or course's web-based resource on at least one occasion (Tab. 8).

A number of prior work evaluated potential privacy harms associated with ed-tech.

Kelly *et al.* [37] aggregate and assess 100 privacy policies from purpose-built ed-tech products used in K-12 schools. The authors use this analysis to build a scoring system for 'Common Sense', a student privacy advocacy organization. However, they do not evaluate products designed for general use.

Notably, CI guides multiple privacy implications analyses of technologies in the educational ecosystem. Rubel and Jones [57] highlight privacy harms associated with learning analytics in higher education. Zeide and Nissenbaum [75] show that data-collection driven information handling practices by Massive Open Online Course (MOOC) platforms and Virtual Education providers violate established norms of traditional education settings. Jones et al. [36] reinforce this notion, showing that learning analytics technologies present "very real challenges to intellectual privacy and contextual integrity" and that "colleges and universities need to make concerted efforts to reestablish normative alignment in concert with student expectations."

Regan and Jesse [56] explore challenges in applying the oft-vaunted "Fair Information Practice Principles" privacy framework [22, 71] to ed-tech. Their concerns center on effects of big-data, autonomy with respect to young people, and surveillance techniques used for purported educational gains. In another work, Regan and Bailey [55] find that education focused journals and magazines largely neglect to cover privacy implications of promoted technologies. Both studies show the contextual specificity of educational privacy concerns and the insufficient governance to meet students' expectations in the context of rapid technological change in education.

Peterson [49] evaluates federal law's failure to adequately protect student privacy in ed-tech, finding that California's attempts to 'band-aid' the gaps highlights the need for overall reform. Given that governance of privacy in the U.S. is highly polycentric, structural governance theory that is compatible with privacy frameworks, such as CI, is useful to understand regulatory and contractual requirements as rules, social expectations as norms, and strategies that bridge gaps and meet local

concerns. The institutional grammar, developed by Crawford and Ostrom [26], and embedded within the governing knowledge commons (GKC) framework [31], is one such approach scholars have employed to study diverse gaps between privacy governance and practice [58].

6 Limitations and Future Work

Future work can expand the survey to include the expectations of students and other relevant stakeholders with remote learning platforms. Our work is also limited to U.S. universities and regulations. We can learn from the remote learning experiences in other jurisdictions, including the effectiveness of alternative regulatory structures. Indeed, we have seen how regulations in Europe affect U.S. institutions by requiring compliance with the General Data Protection Regulation (GDPR) for students who are accessing virtual classes from a location in the European Union [74]. Finally, future work can explore the different governance models at universities for procuring and administrating remote learning platforms, and assess how well they address the concerns of stakeholders.

7 Conclusion and Recommendations

While our results reveal substantial gaps between norms, markets, regulations, and architecture for remote learning platforms, we emphasize that these gaps are not immutable characteristics of the platforms, but rather they reflect issues with default features, settings, and policies to which institutions can negotiate modifications. Our work suggests that DPAs and institutional policies can make platforms modify their default practices that are in tension with institutional values. This approach gives universities the ability to adapt to user expectations—whether for privacy, security, features, usability, or accessibility—and institutionalize these expectations through the negotiation process.

In other words, universities can use their internal policies to bridge gaps between local needs, community expectations, existing regulation, and practice. Accordingly, we recommend that universities use community privacy norms to set the baseline for privacy strategies and practices. By respecting norms and addressing usability concerns, universities can improve the educational experience and reduce the number of instructors who work around supported platforms or defaults.

Crucially, universities do not need to undertake a complex vetting process before licensing software. Instead, we recommend IT administrators establish clear principles for how software should respect the norms of the educational context and require developers to offer products that let them customize the software for that setting. Software developers should commit to use that feedback to continually improve the services. We know that significant user issues surface during software use, especially as platforms' functions or uses creep or are employed in new contexts. The key is to build a process to identify concerns or needs and rapidly fix problems, especially privacy harms, including thwarted expectations, control, and informed choice [21].

We offer the following specific recommendations:

- **Identify User Expectations:** Administrators should periodically solicit concerns and expectations from instructors and students about the major platforms the universities have, or intend to, license. Our survey revealed that instructors were more likely to respond to questions based on specific cases. At the same time, administrators should be sensitive to how certain design choices (e.g., video recordings) may disproportionately impact vulnerable groups who are often targets of online abuse [32,51] and should design surveys to surface such concerns.

- **Negotiate Specific Practices:** While platforms may offer education-specific terms in their contracts, universities should negotiate terms based on an individualized needs assessment. Among the changes universities may want to request are options for local hosting, third-party sharing, limiting how platforms use data, and separating the institutions' data from that of other platform users.

- **Penalize Noncompliance:** Regulators and universities should work together to identify instances of noncompliance and create incentives for the platforms to take prompt action to remediate harms, in the inclusive legal sense [21]. In particular, we recommend strengthening regulations to ensure that software used in educational institutions comply with state and federal laws and that there are mandatory baseline security practices for educational technology that parallel those financial institutions are required to adopt to protect consumer information under the Federal Trade Commission's Safeguards Rule [29].

- **Popularity Does Not Guarantee Security:** If institutions fail to adequately prioritize security, platforms will continue to prioritize growth over product improvements. Our evaluation reflects these misaligned incentives, showing little correlation between product security and popularity.

The shift to virtual learning requires many sacrifices from instructors and students already—we should mitigate their real harms, not further sacrifice usability, security, and privacy.

Acknowledgements

We thank our reviewers for their helpful advice on improving the paper. We are also grateful to members of the NITRD and PLSC for their feedback on drafts of this work.

References

[1] Bugcrowd - BlueJeans. https://bugcrowd.com/bluejeans.

[2] HackerOne - Slack. https://hackerone.com/slack.

[3] HackerOne - Zoom. https://hackerone.com/zoom.

[4] Security Guide: Zoom Video Communications, Inc. Technical report, Zoom Inc. https://web.archive.org/web/20200403154149/https://zoom.us/docs/doc/Zoom-Security-White-Paper.pdf.

[5] Fpf guide to protecting student data under sopipa. Technical report, Future of Privacy Forum, 2016.

[6] Information security and privacy standard terms and conditions. Technical report, Florida State University, 2018.

[7] Multiple Remote Code Execution issues. Technical report, Jitsi, 2020. https://github.com/jitsi/security-advisories/blob/master/advisories/JSA-2020-0001.md.

[8] Selecting and Safely Using Collaboration Services for Telework. Technical report, National Security Agency, 2020.

[9] 34 CFR Part 99 20 U.S. Code §1232g. Family Educational Rights and Privacy Act (FERPA) .

[10] AB-375. California Consumer Privacy Act of 2018.

[11] Martín Abadi, Mihai Budiu, Úlfar Erlingsson, and Jay Ligatti. Control-flow integrity principles, implementations, and applications. *ACM Transactions on Information and System Security (TISSEC)*, 13(1):1–40, 2009.

[12] Mazin Ahmed. Hacking zoom: Uncovering tales of security vulnerabilities in zoom, 2020.

[13] BakerHostetler. State Student Privacy Law Compendium, 2019. https://cdt.org/wp-content/uploads/2016/10/CDT-Stu-Priv-Compendium-FNL.pdf.

[14] David G. Balash, Dongkun Kim, Darika Shaibekova, Rahel A. Fainchtein, Micah Sherr, and Adam J. Aviv. Examining the Examiners: Students' Privacy and Security Perceptions of Online Proctoring Services. *USENIX Symposium on Usable Privacy and Security*, 2021.

[15] Tod Beardsley. Dispelling Zoom Bugbears: What You Need to Know About the Latest Zoom Vulnerabilities. Technical report, Rapid7, 2020. https://blog.rapid7.com/2020/04/02/dispelling-zoom-bugbears-what-you-need-to-know-about-the-latest-zoom-vulnerabilities/.

[16] Ashley Boyd. Which Video Call Apps Can You Trust?, 2020. https://blog.mozilla.org/blog/2020/04/28/which-video-call-apps-can-you-trust/.

[17] Benjamin L Bullough, Anna K Yanchenko, Christopher L Smith, and Joseph R Zipkin. Predicting exploitation of disclosed software vulnerabilities using open-source data. In *Proceedings of the 3rd ACM on International Workshop on Security And Privacy Analytics*, pages 45–53, 2017.

[18] Canvas. What are Conferences?, 2021. https://community.canvaslms.com/t5/Canvas-Basics-Guide/What-are-Conferences/ta-p/53.

[19] Catalin Cimpanu. Zoom to revamp bug bounty program, bring in more security experts. *ZDNet*. https://www.zdnet.com/article/zoom-to-revamp-bug-bounty-program-bring-in-more-security-experts/.

[20] Cisco. Security Vulnerability Policy. https://tools.cisco.com/security/center/resources/security_vulnerability_policy.html.

[21] Danielle Keats Citron and Daniel J Solove. Privacy harms. *Available at SSRN*, 2021.

[22] Federal Trade Commission, Federal Trade Commission, et al. Fair information practice principles. 25, 2007.

[23] Student Privacy Compass. STATE STUDENT PRIVACY LAWS, 2019. http://studentprivacycompass.org/state-laws/.

[24] Connecticut. Zoom guideance, 2020. https://portal.ct.gov/Government/Work-from-Home-Technology-Resources/Zoom-Guidance.

[25] Crispan Cowan, Calton Pu, Dave Maier, Jonathan Walpole, Peat Bakke, Steve Beattie, Aaron Grier, Perry Wagle, Qian Zhang, and Heather Hinton. Stackguard: Automatic adaptive detection and prevention of buffer-overflow attacks. In *USENIX Security Symposium*, volume 98, pages 63–78. San Antonio, TX, 1998.

[26] Sue ES Crawford and Elinor Ostrom. A grammar of institutions. *American political science review*, pages 582–600, 1995.

[27] Cyber Independent Testing Lab. Methodology: "How difficult is it for an attacker to find a new exploit for this software?". Available at: https://cyber-itl.org/about/methodology/.

[28] Electronic Frontier Foundation. Privacy Badger, 2021. https://privacybadger.org.

[29] Federal Trade Commission. Public workshop examining information security for financial institutions and information related to changes to the safeguards rule. *Federal Register*, 85(45):13082–1308, mar 2020.

[30] Susan Freiwald. Comparative institutional analysis in cyberspace: the case of intermediary liability for defamation. *Harv. JL & Tech.*, 14:569, 2000.

[31] Brett M Frischmann, Michael J Madison, and Katherine Jo Strandburg. *Governing knowledge commons*. Oxford University Press, 2014.

[32] Becky Gardiner. "It's a terrible way to go to work:" what 70 million readers' comments on the Guardian revealed about hostility to women and minorities online. *Feminist Media Studies*, 18(4):592–608, 2018.

[33] Michael C Gegick, Laurie Ann Williams, and Mladen A Vouk. Predictive models for identifying software components prone to failure during security attacks. Technical report, North Carolina State University. Dept. of Computer Science, 2008.

[34] Lucinda Gray, Nina Thomas, and Laurie Lewis. Teachers' use of educational technology in us public schools: 2009. first look. nces 2010-040. *National Center for Education Statistics*, 2010.

[35] David Haynes, David Bawden, and Lyn Robinson. A regulatory model for personal data on social networking services in the uk. *International Journal of Information Management*, 36(6):872–882, 2016.

[36] Kyle ML Jones, Andrew Asher, Abigail Goben, Michael R Perry, Dorothea Salo, Kristin A Briney, and M Brooke Robertshaw. "We're being tracked at all times": Student perspectives of their privacy in relation to learning analytics in higher education. *Journal of the Association for Information Science and Technology*, 2020.

[37] G Kelly, J Graham, and B Fitzgerald. State of edtech privacy report. *Common Sense Privacy Evaluation Initiative*, 2018.

[38] Loren Kohnfelder and Praerit Garg. The threats to our products. Technical report, Microsoft, 1999. Internal Microsoft Magazine. Available at https://adam.shostack.org/microsoft/The-Threats-To-Our-Products.docx.

[39] Lawrence Lessig. *Code: And other laws of cyberspace, version 2.0*. Basic Books, 2006.

[40] Chen Ling, Utkucan Balcı, Jeremy Blackburn, and Gianluca Stringhini. A first look at zoombombing. *arXiv preprint arXiv:2009.03822*, 2020.

[41] R. Lock and I. Sommerville. Modelling and analysis of socio-technical system of systems. In *2010 15th IEEE International Conference on Engineering of Complex Computer Systems*, pages 224–232, 2010.

[42] Hector Marco-Gisbert and Ismael Ripoli Ripoli. Address Space Layout Randomization Next Generation. *Applied Sciences*, 2019.

[43] Bill Marczak and John Scott-Railton. Move Fast and Roll Your Own Crypto: A Quick Look at the Confidentiality of Zoom Meetings. *CitizenLab*. https://citizenlab.ca/2020/04/move-fast-roll-your-own-crypto-a-quick-look-at-the-confidentiality-of-zoom-meetings/.

[44] Jennifer Morrison, Steven Ross, Roisin Corcoran, and AJ Reid. Fostering market effciency in k–tech procurement. Technical report, Johns Hopkins University, 2014.

[45] Helen Nissenbaum. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, 2009.

[46] Lindsay Oliver. What You Should Know About Online Tools During the COVID-19 Crisis, 2020. https://www.eff.org/deeplinks/2020/03/what-you-should-know-about-online-tools-during-covid-19-crisis.

[47] oskarsv. Remote Code Execution in Slack desktop apps + bonus, 2020. https://hackerone.com/reports/783877.

[48] PaX Team. Pax address space layout randomization (aslr). 2003.

[49] Dylan Peterson. Edtech and student privacy: California law as a model. *Berkeley Technology Law Journal*, 31(2):961–996, 2016.

[50] J.M. Porup. Bug bounty platforms buy researcher silence, violate labor laws, critics say. *CSO Online*. https://www.csoonline.com/article/3535888/bug-bounty-platforms-buy-researcher-silence-violate-labor-laws-critics-say.html.

[51] Anastasia Powell, Adrian J Scott, and Nicola Henry. Digital harassment and abuse: Experiences of sexuality and gender minority adults. *European Journal of Criminology*, 17(2):199–223, 2020.

[52] Privacy Technical Assistance Center. Protecting Student Privacy While Using Online Educational Services: Requirements and Best Practices, 2014.

[53] Privacy Technical Assistance Center. Responsibilities of third-party service providers under ferpa, 2015.

[54] Qualys SSL Labs. SSL Server Test, 2021. https://www.ssllabs.com/ssltest/.

[55] Priscilla M Regan and Jane Bailey. Big data, privacy and education applications. *Ottawa Faculty of Law Working Paper*, (2019-44), 2019.

[56] Priscilla M Regan and Jolene Jesse. Ethical challenges of edtech, big data and personalized learning: twenty-first century student sorting and tracking. *Ethics and Information Technology*, 21(3):167–179, 2019.

[57] Alan Rubel and Kyle ML Jones. Student privacy in learning analytics: An information ethics perspective. *The information society*, 32(2):143–159, 2016.

[58] Madelyn Sanfilippo, Brett Frischman, and Katherine Strandburg. Privacy as commons: Case evaluation through the governing knowledge commons framework. *Journal of Information Policy*, 8:116–166, 2018.

[59] Madelyn R Sanfilippo, Yan Shvartzshnaider, Irwin Reyes, Helen Nissenbaum, and Serge Egelman. Disaster privacy/privacy disaster. *Journal of the Association for Information Science and Technology*, 71(9):1002–1014, 2020.

[60] Thorsten Schröder. Zoom Endpoint-Security Considerations. Technical report, 2020. https://dev.io/posts/zoomzoo/.

[61] Martin Shudrak. Defeating Windows ASLR via low-entropy shared libraries in 2 hours. Technical report, 2020. https://medium.com/@mxmssh/defeating-windows-aslr-via-32-bit-shared-libraries-in-2-hours-1e225e182155.

[62] Yan Shvartzshnaider, Noah Apthorpe, Nick Feamster, and Helen Nissenbaum. Going against the (appropriate) flow: a contextual integrity approach to privacy policy analysis. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 162–170, 2019.

[63] Yan Shvartzshnaider, Schrasing Tong, Thomas Wies, Paula Kift, Helen Nissenbaum, Lakshminarayanan Subramanian, and Prateek Mittal. Learning privacy expectations by crowdsourcing contextual informational norms. In *HCOMP*, pages 209–218, 2016.

[64] Singer, Natasha and Krolik, Aaron. Online cheating charges upend dartmouth medical school, 2021.

[65] Alexander Sotirov. Bypassing memory protections: The future of exploitation. In *USENIX Security*, 2009.

[66] Jack Tang. Exploring control flow guard in windows 10. Technical report, Trend Micro Threat Solution Teahm, 2015.

[67] Jacob Thompson. Six Facts about Address Space Layout Randomization on Windows. Technical report, FireEye, 2020. https://www.fireeye.com/blog/threat-research/2020/03/six-facts-about-address-space-layout-randomization-on-windows.html.

[68] University of California Office of the President. Electronic Communications Policy, 2005. https://policy.ucop.edu/doc/7000470/ElectronicCommunications.

[69] Luke Taylor Valenta. Measuring and securing cryptographic deployments. 2019.

[70] T. Walshe and A. Simpson. An empirical study of bug bounty programs. In *2020 IEEE 2nd International Workshop on Intelligent Bug Fixing (IBF)*, pages 35–44, 2020.

[71] Willis H Ware. Records, computers and the rights of citizens. 1973.

[72] Martin Weller. Twenty years of edtech. *Educause Review Online*, 53(4):34–48, 2018.

[73] Christian Wressnegger, Fabian Yamaguchi, Alwin Maier, and Konrad Rieck. Twice the bits, twice the trouble: Vulnerabilities induced by migrating to 64-bit platforms. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 541–552, New York, NY, USA, 2016. Association for Computing Machinery.

[74] Gabriela Zanfir-Fortuna. The General Data Protection Regulation: Analysis and Guidance for US Higher Education Institutions. Technical report, Future of Privacy Forum, 2021.

[75] Elana Zeide and Helen Nissenbaum. Learner privacy in moocs and virtual education. *Theory and Research in Education*, 16(3):280–307, 2018.

A Additional Survey Details and Results

Here, we describe our instructor and administrator surveys in detail.

Survey details. The instructor and administrator surveys were hosted on Qualtrics. The surveys were conducted from July 2020 to January 2021. No compensation was given to participants.

The instructor survey begins by asking participants general information about their teaching: their institution, grade level, field, and sizes of classrooms they teach. Next, we ask what video conferencing platforms instructors use (from a multiple choice list), and whether they use a personal or institutional provided version. If instructors use an institution and personal version, they are instructed to mark 'institutional'. We then ask instructors the reasons that instructors choose to use each platform. We then ask instructors the same questions for remote learning platforms other than those used for remote conferencing. For all these questions, instructors have the option to include additional platforms not listed in our survey using an 'Other' option.

Finally, we ask instructors five freeform questions involving their concerns and expectations with remote learning platforms. First, we ask about their own concerns, followed by concerns they have heard from their students. Next, we ask what features instructors consider essential for a remote learning platform to have. Finally, we ask instructors to discuss privacy features and security features that they currently use or desire in a platform.

In the administrator survey, we first ask administrators what remote learning platforms their institution contracted with or supported before COVID-19 in a multiple-choice list. We then ask questions about the administrators' decision making processes for procuring new platforms at their university. We ask whether their universities have a documented process for selecting teaching tools, and whether they would share this document with us. We also ask whether they follow this process strictly, and to explain why if not.

Next, we ask a series of freeform questions regarding platforms at administrators' universities. We ask what remote teaching platforms administrators considered adding since COVID-19 began; what platforms were added since COVID-19; whether admins are trying to request new features or cancel any platform licenses after COVID-19; and what platforms were rejected after review. Administrators are also asked to explain each answer.

Then, we ask administrators to rate the influence of different first parties and third parties on the platform procurement process from 'A great deal' to 'None at all'. We ask how administrators consult these parties (survey, interviews, etc.) We also ask admins to rate the importance of different features of remote learning platforms, from 'Extremely important' to 'Not at all important'.

We ask admins to describe any additional features/protections they negotiated with platforms. We ask whether admins' universities have processes for collecting feedback from instructors/students on platforms currently in use. We also ask which platforms gained users since COVID-19.

We ask admins to rate sources of information by the likelihood of considering them when addressing future platform or policy adoption, from 'Extremely likely' to 'extremely unlikely', and what information would be most helpful for admins to learn. Finally, we allow admins to leave additional comments about remote learning.

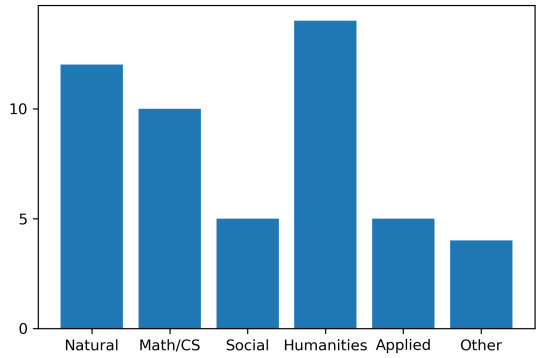


Figure 4: Subjects taught per instructor.

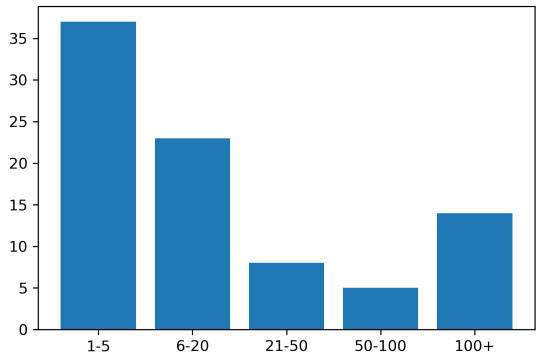


Figure 5: Class video conferencing sizes.

Results. We received 128 total responses to our instructor survey. Of our respondents 109 (85.2%) taught at an undergraduate/graduate level, while the remaining 19 (14.8%) taught K-12 or at professional schools (dentistry, tech literacy, etc.) Respondents came from a diverse set of disciplines and locations, and taught classes from small discussion groups to 100+ lectures.

We asked instructors about their teaching background: the institution they teach at, their class level and sizes, and dis-

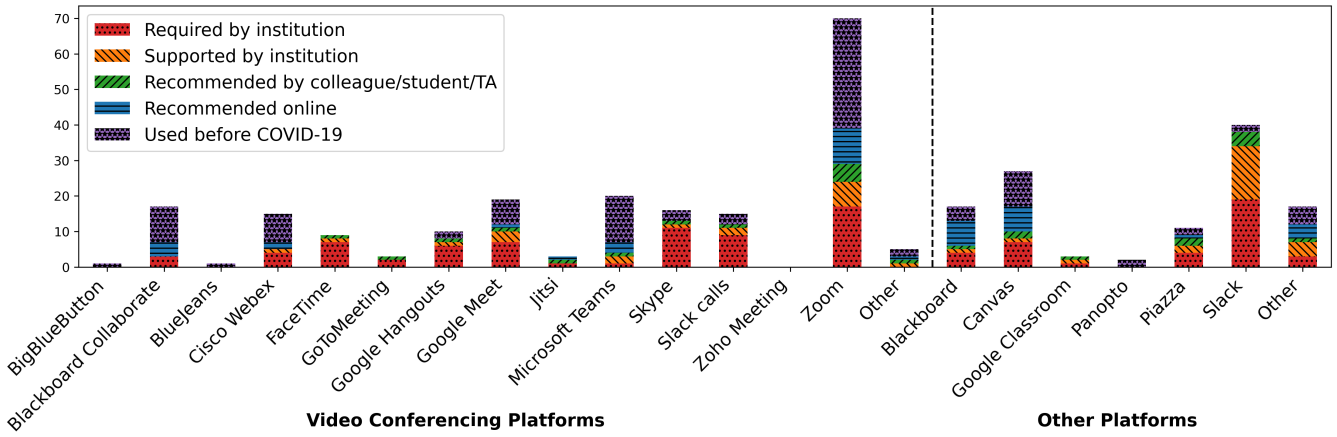


Figure 6: Motivations for instructors using platforms.

cipline area. We then had instructors explain the video conferencing and other remote learning platforms they use, including whether they use a personal or institutional version, as well as their motivations for using each platform. Next, we asked instructors to describe their complaints regarding platforms, and complaints they have received from students and others involved in the course. Finally, we asked instructors what features they consider essential for a remote teaching platform to have—separately for general features, privacy features, and security features.

At least 49 (38%) of instructors were located in the United States and at least 8 (6.2%) in Europe, with others in the Middle East, India, and Australia. The others respondents left this field blank, or gave an abbreviation that matched universities in multiple countries.

Note that we limit all results in the main text and below to undergraduate/graduate instructors from U.S. institutions.

Class subjects and sizes. Figure 4 shows the subjects taught by each instructor. Figure 5 shows the sizes of the groups that each instructor reported regularly videoconferencing with. Note that we allowed instructors to report multiple sizes of groups, so Figure 5 represents the number of conferencing *settings* rather than the number of instructors.

Motivations for platforms. In Figure 6, we present the motivations that instructors reported for using each video conferencing and remote learning platforms. Zoom and Teams were frequently used by instructors prior to COVID-19, while Zoom, Slack and Skype were required by many institutions. These results may highlight the use of tools for communication among other instructors and administrators currently and in the past, in addition to educating students.

Complaints from instructors. Next, we asked instructors to describe their frustrations with the platforms they currently use; 28 (57%) of instructors provided complaints on a wide variety of issues.

Two instructors reported frustration with the fact that no single platform can handle all of their needs, forcing them to "cobble together" multiple platforms to run a course, each with its own learning curve for instructors and for students. instructors cited steep learning curves and/or a lack of documentation for understanding how to efficiently configure and use platforms such as Zoom, Blackboard, and Kaltura.

Other instructors reported frustrations with the fact that they were required to use tools they did not like, with one instructor describing being "forced against [their] will" to use Zoom. In fact, many complaints were directed towards Zoom. Several instructors wished that Zoom would allow students to move between breakout rooms or communicate across breakout rooms to facilitate better class discussions. One instructor disliked that the "Zoom chat history is not available for students who join late," forcing the instructor to repeatedly paste links for tardy students.

Instructors also found it difficult to engage with students, and one wished they could "force students to keep their cameras on to monitor their engagement."

Complaints from students. 24 (48.9%) of instructors also reported complaints from their students about the remote learning platforms they use.

Many students experienced issues with low bandwidth or other glitches that hamper their ability to participate in class. Students felt generally fatigued from conducting class entirely through video conferencing. Students also found it difficult to communicate with other students in the class, and only very small breakout rooms allowed for meaningful engagement.

B Privacy Policy Analysis

In Table 5, we present the breakdown for third-party tracking in platform privacy policies. In Table 6, we present the breakdown for location tracking. A ✓means the policy explicitly permits the activity; an ✗means the policy explicitly forbids the activity; others do not specify.

Blackboard Collaborate supercedes Blackboard's privacy policy in one case, where Blackboard Collaborate does not share data with third-party advertisers (while Blackboard may).

Google's education policy supercedes Google's main primary policy in one case: Google does not share G Suite for Education data with third-party advertisers except for "non-core" products such as Maps and YouTube.

C Android Permissions

When an individual uses an app on either iOS ( ) or Android ( ), the user is promoted to allow or deny various permissions that the app requests. Such permissions vary from the relatively innocuous (take/view photos) to fairly powerful (control system settings). Depending on the permissions model of the operating system the user may be prompted at install time , run time /, or as additional permissions are requested by the app /. Each app's use case requires tailored permissions. However, a developer is free to request permissions that are not essential for the underlying project but instead further the developer's business interests, for example, facilitating the capture of user data for later resale to third parties. Such extra permissions are also inconsistent with the security principle of least privilege to which mobile OS developers attempt to adhere: no app should have more permissions than it requires, to do otherwise increases the attack surface area. Thus, while the presence of any particular permission granted to an app may not on its own be reason for concern, over-broad requests for permissions pose problems for both security and privacy.

Results. We collected the set of requested permissions for the Android versions of the products from each app's page on the Google Play Store website. As iOS apps obtain permissions when the associated action is attempted (rather than at install time, as on Android), gathering the equivalent data for the operating system would require exhaustively interacting with all app features, a task we leave to future work. We tabulate the full set of results in Table 7.

We found that the applications requested between fourteen and twenty-eight different permissions, with the average application requesting twenty-three permissions. More or fewer permissions requested is not inherently better or worse. However, apps with more permissions do carry a higher risk of violating the principle of least privilege, and therefore of facilitating privacy and security violations.

D Network Traffic Domains

We present the list of first-party and third-party domains contacted by each desktop platform during our network traffic analysis, along with the type of service provided by each third party, in Table 8. The Jitsi client only contacts the domain that a user specifies, so we do not report any contacted domains for it.

E Binary Security Feature Descriptions

SafeSEH (Windows Only). Safe Structured Exception Handling (SafeSEH) is a mechanism to ensure that only authorized exception handlers execute [65]. A binary with SafeSEH contains a list of exception handlers, which the kernel stores in a protected list when the program begins execution. If an exception is thrown, the kernel checks if the handler is pre-approved and, if so, allows the handler to execute.

DEP/NX. Data Execution Prevention techniques separate areas of memory that contain data and those that contain code, restricting the user from executing code contained in areas marked for data. No Execute bit (NX) is a CPU implementation of the DEP concept. A binary with support for DEP/NX has appropriately marked memory regions.

ASLR. Address Space Layout Randomization (ASLR) randomizes the layout of a binary when the operating system loads it into a virtual address space [48]. Its efficacy is tied to the number of possible different layouts. 32-bit operating systems traditionally reserved only randomized 8 bits, giving an attacker a 1/256 chance of guessing the layout [42]. Modern 64-bit versions of MacOS and Windows support randomizing up to 16 and 19 bits of entropy respectively (corresponding to 66k and 524k guesses). Notably, Windows requires two compile time flags to be set to enable 19 bit ASLR, without which at most 14 bits can be randomized (16k guesses) [61, 67].

CFI. Control flow integrity refers to a class of mitigations that aim to prevent an attacker from redirecting program flow [11]. Microsoft's implementation of this concept, Control Flow Guard (CFG), adds a check before `call` instructions (that transfer execution to a function) that do not have static arguments. The check cross-references a data structure that stores the start address of all valid functions, and throws an exception if the program execution would otherwise be transferred to an invalid address [66].

Code Signing. Code signing provides a chain-of-trust to validate authorship of a binary. Both Microsoft and Apple provide services for third-party developers to sign their applications. Both prevent users of their most modern operating system versions from executing unsigned binaries absent explicit acknowledgment.

Stack Canaries. Stack canaries are a compile-time modification that allow a program to detect a subset of buffer overflow attacks [25]. A canary value is placed on the stack between a stack frame's return pointer and other variables. Before a program uses the return pointer on the bottom of the stack, it first checks the contents of the stack canary against a known value. If the value has been altered, the stack has been tampered with, and the program will terminate with an error.

Architecture Width. While the architecture for which a binary is compiled is not inherently a security feature, 32-bit

Third-Party Tracking	Burden on Users to Monitor	Shared With Advertisers	Bi-directional Sharing	Social Media Data
Apple		✗		
BigBlueButton	✓			✓
Blackboard		✓		✗
Blackboard Collaborate		✗		✗
BlueJeans	✓	✓	✓	✓
Canvas		✓		✓
Cisco				✓
Google		✓		✓
Google Education		✗*		✓
GoToMeeting/LogMeIn	✓	✓	✓	✗
Jitsi		✗		
MS Teams/Skype	✓		✓	
Panopto	✓		✓	
Piazza	✓	✓		
Skype for Business				
Slack	✓	✓	✓	
Zoom		✗		
Zoho	✓	✓	✓	✓

Table 5: The breakdown of third-party data sharing in privacy policies. * = Google's G Suite Education policy specifies that data is only shared with third parties for non-core services, such as Maps and YouTube.

Location Tracking	Permitted	Active Tracking	Shared with Third Parties	Inferred Location	Exact Location
Apple	✓			✓	
BigBlueButton	✗	✗	✗	✗	✗
Blackboard	✗	✗	✗	✗	✗
Blackboard Collaborate	✗	✗	✗	✗	✗
BlueJeans	✓		✓		✓
Canvas		✗	✗	✗	
Cisco	✓	✓	✗	✓	✗
Google	✓		✗	✓	✓
Google Education	✓		✗	✓	✓
GoToMeeting/LogMeIn	✓		✓		✓
Jitsi	✗	✗	✗	✗	
MS Teams/Skype	✓			✓	✓
Panopto	✓		✓		✓
Piazza	✓			✗	✓
Skype for Business	✓	✓	✓	✗	✓
Slack	✓	✗		✓	✓
Zoom	✗	✗	✗	✗	
Zoho	✓	✓	✗		✓

Table 6: The breakdown of location data in privacy policies. Active tracking means A-GPS or WiFi location tracking on a continuous basis. Inferred location data includes IP address-based locations and other inferences.

binaries are unable to use many modern OS security measures. To this end, MacOS 15 (Oct 2019) ceased support for 32-bit binaries. We therefore checked that the Windows software packages were 64-bit. (Note that 32-bit software packages that developers migrate to 64-bit architectures may exhibit pathologies [73]. Thus, packages that have recently transitioned to 64-bit architectures merit further caution.)

Permission	BlueJeans	Jitsi	Slack	Teams	WebEx (M)	WebEx (T)	Zoom	Totals
Version	41.1813	20.2.3	206.10	1416	40.6.1	4.11.241	5.1.27838	
access Bluetooth settings	✓	✗	✗	✓	✗	✗	✓	3
access download manager	✗	✗	✗	✓	✗	✗	✗	1
add/modify calendar events and send email...	✗	✓	✗	✗	✗	✗	✓	2
add or remove accounts	✓	✗	✗	✓	✓	✓	✓	5
approximate location (network-based)	✓	✗	✗	✓	✓	✓	✓	5
change network connectivity	✗	✗	✗	✓	✗	✗	✗	1
change your audio settings	✓	✓	✓	✓	✓	✓	✓	7
connect and disconnect from Wi-Fi	✗	✗	✗	✗	✗	✓	✗	1
control vibration	✓	✗	✓	✓	✗	✓	✓	5
create accounts and set passwords	✗	✗	✗	✓	✓	✓	✗	3
directly call phone numbers	✓	✗	✗	✓	✓	✓	✓	5
disable your screen lock	✓	✗	✗	✗	✗	✗	✗	1
download files without notification	✗	✗	✗	✓	✗	✗	✗	1
draw over other apps	✓	✓	✗	✓	✓	✓	✓	6
expand/collapse status bar	✗	✗	✗	✓	✗	✗	✗	1
find accounts on the device	✓	✗	✓	✓	✓	✓	✓	6
full network access	✓	✓	✓	✓	✓	✓	✓	7
install shortcuts	✗	✗	✗	✗	✓	✗	✗	1
modify or delete USB storage	✓	✓	✓	✓	✓	✓	✓	7
modify system settings	✗	✗	✗	✗	✗	✗	✓	1
modify your contacts	✗	✗	✗	✓	✗	✓	✗	2
pair with Bluetooth devices	✓	✓	✓	✓	✓	✓	✓	7
precise location (GPS & network-based)	✓	✗	✗	✓	✓	✗	✓	4
prevent device from sleeping	✓	✓	✓	✓	✓	✓	✓	7
read calendar events plus confidential info...	✓	✓	✗	✗	✓	✓	✓	5
read phone status and identity	✓	✓	✓	✗	✓	✓	✓	6
read sync settings	✗	✗	✗	✗	✓	✗	✗	1
read USB storage	✓	✓	✓	✓	✓	✓	✓	7
read your contacts	✓	✗	✓	✓	✓	✓	✓	6
receive data from Internet	✓	✓	✓	✓	✓	✓	✓	7
record audio	✓	✓	✓	✓	✓	✓	✓	7
reorder running apps	✗	✗	✗	✗	✓	✗	✗	1
retrieve running apps	✗	✗	✗	✗	✓	✗	✗	1
run at startup	✓	✗	✓	✓	✗	✗	✗	3
send sticky broadcast	✓	✗	✗	✗	✓	✗	✓	3
take pictures and videos	✓	✓	✗	✓	✓	✓	✓	6
toggle sync on and off	✗	✗	✗	✗	✓	✗	✗	1
uninstall shortcuts	✗	✗	✗	✗	✓	✗	✗	1
use accounts on the device	✓	✗	✗	✓	✗	✗	✓	3
view Wi-Fi connections	✓	✓	✗	✓	✓	✓	✓	6
view network connections	✓	✓	✓	✓	✓	✓	✓	7
Totals	26	15	14	28	28	23	26	138

Table 7: **Permissions Requested by Android Apps.** We tabulated the permissions requested by the different applications when installed on Android. We sourced permissions from the app listings on Google Play.

Domains Contacted					
BlueJeans	**Jitsi***	**MS Teams**	**Slack**	**WebEx**	**Zoom**
bluejeans.com **hockeyapp.net** (Analytics) **mixpanel.com** (Analytics) **nr-data.net** (Analytics)		microsoft.com msedge.com	**chime.aws.com** (Video conferencing) **gravatar.com** (Graphic avatars) slack.com slack-edge.com slack-ims.com	webex.com	zoom.us

Table 8: The domains contacted by each platform in our network analysis. **Blue** domains are third-party domains. * = Jitsi requires specifying a server domain to connect to, and does not connect to other domains.

Challenges and Threats of Mass Telecommuting: A Qualitative Study of Workers

Borke Obada-Obieh Yue Huang Konstantin Beznosov

University of British Columbia
Vancouver, Canada
{borke,huang13i,beznosov}@ece.ubc.ca

Abstract

This paper reports the security and privacy challenges and threats that people experience while working from home. We conducted semi-structured interviews with 24 participants working from home in the three weeks preceding the study. We asked questions related to participants' challenges with telecommuting. Our results suggest that participants experienced challenges, threats, and potential outcomes of threats associated with the technological, human, organizational, and environmental dimensions. We also discovered two threat models: one in which the employer's asset is at stake and another in which the employee's privacy is compromised. We believe these insights can lead to better support for employees and possibly reduce cyber-attacks associated with telecommuting during the pandemic and beyond.

1 Introduction

Our research aims to provide insight into the security and privacy concerns associated with telecommuting to help employees safely work from home while protecting organizations' confidential information. Our investigation into telecommuting challenges is a response to a clear need for safer work-from-home practices as the rise in telecommuting has led to an increase in cyber-attacks [6, 31, 37, 44].

The global COVID-19 pandemic has resulted in the world's largest telecommuting situation [5]. In 2018, the U.S. Bureau of Labor Statistics report showed that only 8% of all employees work from home at least one day of the week, while 2% worked fully from home [7, 45]. However, most employees now work from home. Recent research from Stanford indicates that as of June 2020, 42% of the labor force was telecommuting (with 33% unemployed and 26% working in essential services) [8, 78]. Researchers estimate that employers plan to keep 20% of their workers continue working from home after the pandemic ends, mainly to reduce costs [41]. Another recent survey shows that 47% of the respondents aim for their workers to telecommute full-time [24]. Further, some major tech companies have already switched to either long-term or permanent work-from-home model [15, 16, 32, 55].

With a remote workforce and everyone working digitally, the threat landscape increases. Research shows that 91% of respondents experienced an increase in cyber-attacks as a result of employees telecommuting [6]. Further, the Canadian Press reported a 1,350% increase in cloud-related attacks and a 4,000% increase in ransomware emails [40]. Remote working can also be problematic when employees' personal computers are not updated with the most recent security protocols and software. Employees risk exposing the entire system to various types of cyber-attacks. Major organizations have suffered data breaches targeted at employees. For instance, the World Health Organization reported a fivefold increase in cyber-attacks, with the most recent attack targeting their employees [77]. There has been a spike in phishing attacks in Italy as a result of people teleworking [33]. In addition, the threat model in a home environment differs from that seen in the physical office workplace. For instance, some company devices used in teleworking are linked to home or less secure Wi-Fi networks. These company devices may not have the physical security provided in the workplace.

To address the security and privacy concerns of working from home, research is needed to understand the specific challenges and threats that employees experience while telecommuting. Several telecommuting research projects compared workers' productivity while working from home and in physical office locations [2, 4, 9, 51, 60]. Some research on working from home also provides tips and strategies for securing the home internet network for employees while telecommuting [21, 35, 43]. However, to the best of our knowledge, no

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

research has focused on employees' security and privacy concerns with telecommuting.

To this end, our objective is to address the following research question: *What are employees' security and privacy challenges, threats, and perceived outcomes of threats when working from home?*

For the sake of clarity, we define some terms. Challenges and threats are often used interchangeably; however, they do not necessarily mean the same thing. In this paper, we define a threat as "an event or condition that has the potential for causing asset loss and the undesirable consequences or impact from such loss" [59]. A challenge is a circumstance that could lead to a threat. And an outcome of a threat is "an expectation of loss expressed as the probability that a particular threat will exploit a particular vulnerability with a particular harmful result" [64]. These outcomes could be loss of organizational data confidentiality, integrity, and availability or the loss of personal privacy. We define confidentiality as "the property of non-public information remaining accessible only to authorized parties" [72]. Privacy "more narrowly involves *personally sensitive* information, protecting it, and controlling how it is shared. ... What information should be private is often a personal choice, depending on what an individual desires to selectively release." Integrity is defined as "the property of data, software or hardware remaining unaltered, except by authorized parties" [72].

We addressed our research question by conducting semi-structured interviews with 24 participants. They were employees who had been working from home in the three weeks preceding the study. We asked questions relating to their challenges with telecommuting and analyzed the results using thematic analysis.

Our study makes two major contributions. First, we performed the first qualitative study on employee security and privacy concerns when telecommuting. Furthermore, we identified the perceived outcomes of threats associated with these concerns. We grouped our findings into four categories of challenges and threats: technological, human, organizational, and environmental. We further grouped our findings into the identified outcomes of threats to security and privacy and created threat models that emerged from our results.

Second, we discovered concerns that need to be addressed to protect employee privacy while telecommuting. Many employees felt that they had to sacrifice some privacy to get their work done, such as revealing their personal phone number or street address to clients. Participants feared that clients could locate their home or that they could suffer a break-in. Therefore, there is a need for discussion of how employees and organizations can protect their privacy and security while telecommuting.

Our contributions provide insights into the security and privacy gaps that exist with regard to employees telecommuting. We are optimistic that these insights can lead to changes in the way telecommuting is currently being carried out. These changes will be helpful during the current pandemic and in other situations where employees need to telecommute.

2 Related Work

Telecommuting and telework are similar but different. Whaley [75] defines telecommuting as "using information and communications technologies (ICTs) to bring work to the worker, rather than require them to go to the work." In telecommuting, the employee does not commute to get to work. Examples of telecommuting could be working in the home office or working out of the office in the home environment, for example, the guest house. On the other hand, telework refers to work that is done somewhere that is a distance from one's office. Examples of teleworking could be working at another branch of an office or working at a telework center with other colleagues. While some types of telework are telecommuting, not all types of telecommuting are telework [58, 75].

Many previous papers focused on teleworking benefits and aimed to understand problems that stop its widespread adoption by organizations. For instance, Pyöriä [53] conducted a literature review on the advantages of distributed work, which the author refers to as telework. Similarly, Kintner [34] conducted surveys with 1,002 respondents to determine how receptive businesses were to telework and identify ways to encourage managers to telework. The respondents were workers in various organizations who were not teleworkers. The author identified issues that prevented telework adoption, such as inadequate security for protecting transmitted information while teleworking, the high cost of buying the needed equipment, and the lack of staff available to aid telework transition, among others. Our study builds on previous research and conducts qualitative research with telecommuters. We chose to interview telecommuters to understand their security and privacy challenges and threats when working from home.

Some papers explored the reasons behind the low adoption of telework before the pandemic. One of the reasons for low adoption was poor data security. Clear and Dickson [17] for instance, studied whether telework adoption was influenced more by levels of worker autonomy, employment flexibility, and management attitudes than technology provision. The authors conducted 303 surveys and 58 interviews with representatives of small and medium enterprises (SMEs). In discussing their results, the authors remarked that data security is "a major disadvantage to the adoption of telework." However, the authors did not explain why this was the case. Spinellis et al. [66] also hypothesized that SMEs lacked the potential to have good technical expertise to maintain an adequate security level in teleworking. The work of Pyöriä [52] is closest to ours. This author conducted a survey and interviews with employees to understand the low adoption of telework even in big organizations. The participants, however, were not teleworkers. The authors categorized their findings into those relating to the individual, the organization, and the community. They

described the pros and cons of telework at each level. The findings relating to the organization level are closest to our findings. The author found that some of the drawbacks of teleworking include the problem of employers seeking new means to surveil and control employees, poor data security, and disruption of privacy in employees' homes. Our work differs from Pyöriä's [52] in two major ways. First, we interview employees who are currently telecommuting. Our focus on telecommuting employees helped us to understand the specific challenges these people are facing. Further, building on Pyöriä's study, we focus on telecommuters' security and privacy concerns and find more challenges and threats. Because of the potential for a number of telecommuters to continue for the long term, our research becomes even more critical.

Several papers focus on the security and privacy challenges of telecommuting. However, these papers are not based on empirical data but on hypothetical situations. For instance, one of the earliest papers on telecommuting was written by Sturgeon [67]. The author used a hypothetical case study to highlight vulnerabilities. The author predicted threats and risks to organizations' confidential data when telecommuting using the Simplified Threat and Risk Assessment Process [67]. A more recent paper by Okereafor and Manny [46] provides an overview of security issues that are related to telecommuting and videoconferencing apps. The authors predicted issues related to workers' geographic location such as workers' telecommuting in locations with poor Wi-Fi networks and workers being distracted while working from home, which could lead to dangerous errors. The authors also highlighted other general issues such as telecommuting devices using a lot of bandwidth and reduction in employees' productivity while working from home.

Our paper is the first to provide a qualitative study on telecommuters to understand their security and privacy challenges, threats, and perceived outcomes of threats. We chose to conduct a qualitative study to understand *why* people face some of the predicted challenges and *how* they experience them. Qualitative studies help answer "why" questions and provide an in-depth understanding of what is being studied [54]. We believe that a more in-depth analysis of these concerns will help researchers better understand the challenges and start a discourse on the ways of addressing them.

3 Methods

3.1 Participant Recruitment

We recruited participants by advertising on Facebook, LinkedIn, and Kijiji using the platforms' paid advertisement functionalities. Potential participants filled out an eligibility survey. To be eligible to take part in the study, participants had to be 19 years or older. Participants had to have worked full-time physically in an office space in the year preceding the study. Participants had to have been working with computers for at least three days a week, so that we could explore current challenges they might be facing with the technology. Further, participants had to have been working remotely full-time in the last three weeks preceding the study. The latter inclusion criteria was to ensure that participants would remember recent experiences with working from home.

3.2 Interview Procedure

We proceeded with the interviews after the participants gave informed consent to participate in the study. To avoid priming, we told participants that the aim of the study was to understand their experiences working from home.

We asked participants for demographic information and about their general experiences working from home. Based on these experiences, participants were asked further questions regarding what they enjoyed about working from home and what they would love to change about their experience (if anything). Participants were also asked to list new technologies that they had been using to work from home. We asked further questions about participants' thoughts about using the technologies (see appendix D). Afterward, we compensated the participants. One or two researchers took part in each interview session. All interview sessions were audio recorded.

3.3 Data Collection

We piloted our study procedure with two participants. Based on the feedback from the pilot interviews, we improved the clarity of the questions. All other instruments in the main study remained the same as those used in the pilot.

We carried out semi-structured individual interviews with all recruited participants. This allowed them to express their thoughts in their own way and to add information as they saw fit, without the restrictions of structured interviews [19].

All interviews were conducted either via Skype or Zoom, based on participants' choice. We chose to conduct online interviews due to the restrictions placed on in-person meetings resulting from the COVID-19 pandemic. Participants were compensated with CAD $20, sent via e-transfer. Data collection was done from March to September 2020. Our university's Behavioural Research Ethics Board (ID: H20-01219) approved the research before any data collection took place.

3.4 Data Analysis

Two researchers transcribed and coded more than 18 hours of recorded interview sessions, each an average of 44 minutes long. Interviews were analyzed using thematic analysis [27], a "set of procedures designed to identify and examine themes from textual data in a way that is transparent and credible" [26]. We followed the data analysis steps outlined

by Guest et al. [26]. Two researchers segmented and coded the transcribed interviews into categories, types, and relationships to develop the codebook. Afterward, three researchers identified the themes that emerged from the data. In addition, four researchers engaged in a code and theme sorting exercise to come to a consensus on the identified themes. We conducted data analysis concurrently with the collection and reached theoretical saturation after 21 interviews, as no new codes emerged from the last three data collection sessions (see saturation graph in Figure C.1).

4 Results

We present our findings in the form of the challenges and perceived threats, which we categorized into technological, human, organizational, and environmental dimensions. We also link them to perceived outcomes of threats.

4.1 Participants

We recruited a diverse set of 24 Canadian participants. They were 19 to 64 years old (mean 41 and median 38), with 14 of them identified as men. Table A.1 shows the demographics of the participants regarding age, gender, educational level, place of work and job, as well as size of the employer and geographic region (when available).

4.2 Technological Dimension

Challenges related to technological dimensions are due to the use of technology while telecommuting. These challenges could result in threats to the security and personal privacy.

4.2.1 Sharing work information in unauthorized ways

Some participants used unofficial online communication channels to share work-related information. This action was a security concern as it was unclear whether these unauthorized technological solutions satisfied employers' data security requirements. Since different communication solutions have varying degrees of compliance with organizations' security and privacy requirements, using these solutions could lead to various security and privacy threats for both the organization and its employees. This action could also lead to the outcome of threat of the loss of the data *confidentiality*.

One reason for using unauthorized channels was **low usability of the authorized channels**. For instance, P15 (customer service representative) was supposed to use Bell Total Connect (BETC). However, he found it unusable: *"[To use BETC] you've got to request access, then you download it, and then you've got to have your credentials in place. ... It's a complicated program."* P15 ended up using Facebook and sometimes text messages to communicate work-related information with his boss and colleagues while telecommuting.

Another reason for the use of alternative communication channels was because **most of our participants' colleagues were already using them.** It was therefore easier to reach colleagues there. P14 (call center representative) explained: *"We do have a chat [function] in our [official] program, [but it's] just that everybody's on Facebook Messenger. So whether you like Facebook or not, you're kind of forced to use Facebook. And so I [use Facebook since] everybody's there."*

4.2.2 Sacrificing personal privacy and security

There were many instances where participants sacrificed their privacy or security to telecommute. We discuss these instances below.

The tension between professionalism and privacy on video calls. Many participants experienced tension and uncertainty around the use of their webcams during work meetings. For the sake of personal privacy, participants wanted to keep their video cameras off during some periods of work calls. However, they were uncertain whether doing so made them appear less professional or serious about their job. For P16 (planning department director), having the webcam on during work meetings was a necessity, although his colleagues did not necessarily agree: *"People should be available on video if they're doing work during the workday. [However,] that [is] a concern for some people. I have a colleague, and today she said, 'I can't show you my video because my hair is in an Afro.' ... Maybe she didn't want people to say something, or to notice, or to make a case out of it."*

P21 (senior project manager) also explained the dilemma: *"I can't force [people to turn their video on]. It's their home, so I can't really force them; I can only insist. I know that some of the managers in our organization make [a] point of telling [employees to] turn [their video] on during the meeting, [because the employees] have to be paying attention."*

Some participants felt that having the video camera on was an invasion of their privacy. Participants feared that people could take screenshots of them without their consent. P18 (executive director), explained this concern: *"I've thought about [people taking screenshots during video meetings], 'cause I know people who have [done that]. I have a call every two weeks, and there's usually about eight or nine of us [on the call], and I know that they're taking screenshots of the video [meeting], but I wish ... a part of me feels like, there should be a notification feature [on the teleconferencing app that shows] if somebody's doing a screenshot [during meetings] or if they save an image. My preference is that people ask if they're going to do a screenshot for whatever reason."* Having webcams on also virtually invited co-workers into participants' homes, which was seen as a privacy invasion. P5 (sales director) explained: *"[Through video calls,] you're inviting a lot of people into [your] home that [you] wouldn't have otherwise. So you're here [on the video call], your kids are walking by, or other family members or your dog or whatever the case*

may be, [and] you may not want people to see [all of that]." P3 (research assistant) further explained: *"[Work video meetings] certainly blur that line between your home life and your workplace. Like right now, you're in my kitchen with me. Normally co-workers wouldn't necessarily be inside the house, which is sort of a weird ... it changes that relationship [with my co-workers].*" Having webcams on during work meetings leads to the loss of employee's *privacy*.

While some of these challenges can be solved using virtual backgrounds [65, 68, 80], participants had issues with the availability and usefulness of virtual backgrounds. First, not all videoconferencing apps fully support virtual backgrounds [39]. Second, not all participants liked the idea of using a virtual background as they found the concept of virtual backgrounds to be too dull or unexciting. Third, virtual backgrounds do not guarantee that people walking by will not pop up on the screen [57].

The design of some tools made it difficult for employees to maintain security while working from home. This challenge sometimes led to the organizational outcome of threat of the loss of the data *confidentiality*. For instance, phones that used the same port for charging and connecting headphones were a challenge in case of long and frequent calls: *"I think the biggest issue [with working from home] for me is [my phone]. If I've got a day that is heavily focused on a lot of client stuff, then I have to continue using my work phone, which can be problematic ever since they've got rid of the bloody plugin that you can put your headphones in and [replaced] it with [one port], because that's [the port] I need to charge my damn phone with. So I have, on occasion, had it plugged in [to charge] and used it without headphones. And technically, depending on the voice tones of the other person [on the other end of the phone], somebody may have [over]heard our conversation.*" [P11 (health director)]. This was a security concern because housemates could overhear confidential information (§4.3.1). The participant sacrificed the confidentiality of his work calls to get the job done. In some cases, to use headphones and maintain security, P11 switched from taking calls on his work phone and used his personal phone instead. However, our results also suggest that using personal phones to manage work conversations could be a security and privacy concern, as we explain below.

Employees share their personal information to aid telecommuting. Some participants shared personal phone numbers or home addresses with colleagues and clients. In some cases they used their personal devices to work from home. These actions sometimes made it difficult for participants to draw the line between their personal and work lives. P8 (accounting supervisor), for instance, could not "move" his work landline home. So he gave the clients his personal phone number. Prior to working from home, P8 never picked up calls from unknown numbers because he was afraid of being scammed. That changed after giving his personal number to work clients: *"[Recently I received a call from an unknown number.] First ... I wasn't going to answer [but] then I [decided to] answer [and] I was really lucky that I took the call because it was [from] the government. And [the government] was just verifying information so that they could pay [my organization] the subsidy. So if I'd refused that call, it would have really slowed down the payment, and then my boss would have been mad at me, because we were rushing around to submit our application. So of course now I'm answering more calls on my [personal phone], and I don't screen it as closely as I [used to do] before. If I'm going to work from home, that's part of working from home. I'm going to pick up the phone for numbers that I don't know.*" P8 sacrificed his privacy and precautionary safety measure to continue his regular work activities at home.

When asked if he still had a fear of picking up a call from a scammer, P8 replied: *"I'm afraid if I pick up [a] call from a scammer, that somehow they are going to know that there is a live person at the end of the line and then they're going to get me more scam calls. [But] I'm afraid that if I miss a business call, then I'm going to get criticized by my boss because it affects my work, [and] I [end up not] do[ing] something [at work] fast enough. And the boss will be mad, because I didn't pick up a phone call.*"

P23 (school secretary) further remarked: *"I had to use my own personal cell phone to communicate with parents. That part of [telecommuting] was awkward ... because now I find the parents text me or leave me a message to get information. For me [giving out my phone number] does cross the boundary. I always have tried to separate as much as I could, my private life from my work life ... it was basically just assumed upon us [by the organization] when [the organization] decided they were going to [send us home to work]. ... I probably could have done [the call blocking code], but I didn't do that. I do believe you get charged for [doing that] so I didn't want to have that fee on top of other fees.*"

In some cases this challenge included giving coworkers participants' home addresses. P3 further explained: *"So if I asked my coworker to pick something up from my office, then probably he might drop it off at my house. So then he would know where I live. So I feel like it starts to open up some kind of personal privacy [issue].*"

The use of some technological tools in telecommuting made it easy to monitor participants' activities. For example, the User Presence feature [70] in Microsoft Teams makes it easy to determine a user's activities online. Some participants were concerned of their *privacy* being further reduced by this feature, as illustrated by P16: *"I notice that you can tell who is on their computer and who is not, [using Microsoft Teams]. For example, now I can type any name, and I can see [who is online and who is not]. [The] red [button] means that they're on a [Microsoft Teams] call or [in] a meeting; green means that they're on their computer, but not in a meeting. And yellow means that they've walked away from their computer and the little X means the computer's turned off. I*

find that [that] can be used to monitor whether people are at their desk or not. So, for example, a manager can check whether their employee is yellow, green, or red, and they could be green and surfing the 'net, and they could be yellow and reading a document [on] the computer. ... [Managers] might jump to conclusions [in] thinking that an employee should be either green or red, but not yellow, because yellow means that they're not [at] the computer."

Unauthorized people controlling participants' computer remotely. The possibility that people's computers could be remotely controlled was a privacy and security challenge for participants. Some jobs require participants to give their employers or customers remote access to their computers. However, in giving employers remote access, participants feared that their employer would be able to access other parts of their computers remotely which could lead to *unauthorized access to data* and loss of *privacy*. When teaching students online, the job of P17 (education assistant) requires her to give her students remote access to her computer so that they can play an educational game: *"When we are sharing the screen with [another] person, we ... give [remote] control to the other person, [and] that was [a] concern because that person can go on your computer and probably check anything on your desktop. [For example, after giving remote control to a student], then that student can control my screen ... or can check anything."*

4.2.3 Reducing security for usability

To make some technological tools usable, security was sometimes sacrificed. We discuss some instances where security and privacy were sacrificed for usability while telecommuting.

Employees bypassed organizations' security measures to make use of technological tools. As a security measure, some work-from-home phones were too locked down, and participants did not find them usable enough. Participants sometimes came up with workaround solutions that were less secure. These workarounds would result in even higher consequence of threat to the *confidentiality* of the organization's data than the task they were trying to accomplish, as illustrated by the story of P6 (senior staff): *"The [work] iPhone that I [use] is so well locked down that I cannot copy and paste from an email into a text message. [If I try to do that, the work iPhone] says 'You cannot paste your organization's data here,' and it's a complete pain because there are times when [I'm] communicating with my boss by text message where she says, 'Can you just send me that phone number?' [or] like an email address or something like that. [I] can just type [the information my boss is asking], but my memory is terrible. I would always copy and paste something rather than [type] it. [It's] a particularly annoying feature and so I found a workaround: If I had something that I needed to text to my boss, I [would] actually send the email from my work email address to my home email address, then use my [personal] iPhone to cut and paste the information into a text and send."*

Reduced security of technology to aid usability. To enable employees to work effectively from home, sometimes IT personnel reduced the security of some organizations' devices. Such compromises could reduce organizational data *confidentiality* and *integrity* and violates organization's security control rules and policies. P8 narrated a related experience: *"[I] brought [a second] monitor home when I first remote accessed [in to work]. The second monitor did not work, and so I complained to the IT manager, and [the IT manager] said [that] for security purposes the standard remote logging software simply does not allow two monitors. So the IT manager said, '[P8's name], don't tell anybody else this because it's not good control, but I made you a special URL, and now you can access [the work computers remotely with] two monitors.' I'm guessing that by giving me this special URL [designed just] for me, I have more access to the [organization's] information... . So I think it's weaker control over the security of [people's] information.' And [the IT guy] did tell me, 'Don't tell anyone else; I'm just doing this as a favor for you,' because IT [has] to maintain the security of the computer network. And if there was a hack or break-in, [the IT manager] would get blame[d]. So I have not told anyone else, but really I should tell my colleagues because it would speed up their work, [but] I'm afraid I'll lose the special favor with the IT manager if I tell anybody else."*

4.3 Environmental Dimension

There were threats specific to the home work environment. They were mostly expressed as fears and concerns. We describe these threats below.

4.3.1 Household members can access the organization's confidential information

There were concerns about others in the household overhearing the organization's confidential information. This was a particular concern for participants with housemates. In some cases, participants shared office space with their housemates. In other instances, the house had thin walls, and the house occupants and guests could overhear conversations held in various locations within the house. Some participants' jobs included handling confidential information; therefore, a security threat was that others could overhear these conversations. This led to the organizational outcome of threat of the loss of data *confidentiality*. For instance, P15, who had three roommates and worked from the dining area of his house, explained: *"If [clients are] giving me [their] credit card information, and I'm reading [the credit card details] back to [them, I would be] around people [in the house while reading the details]. Frankly, I don't think I'll be able to avoid [my roommates' overhearing] until I go back to the office. ... Right now, if somebody comes into the kitchen [to] make food, I could be*

on a call, [and] that makes things a little awkward at times."

Participants feared that their customers and colleagues could overhear private conversations from participants' homes. They were concerned with the loss of their and other housemates' *privacy*. P9 (call center agent), explained this concern: *"We have very thin walls in my house, and my room is right beside the bathroom. And a lot of times when my parents are calling [for] my brothers' [attention], I can hear [my parents] through the wall. Sometimes I have text[ed] my brothers [saying], 'Hey, can you please keep it down? I'm on the phone with a taxpayer. And they may be able to hear you through my headset.' [At] home you can almost hear everything that goes on."*

There was also the possibility of unauthorized people viewing employer's confidential data. For instance, P11, who worked from his dining room, explained: *"[I] had multiple eye surgeries last year, so I don't really see out of this [eye]. So I have a big screen in our dining room, which is completely open to our kitchen. And then [on] another side, it's kind of an open concept: living room, dining room, [and] kitchen. If anybody was coming in and walking around, they could have seen documents that I was working on the large screen, because it blows it up quite large, so it's quite legible to anybody that wanted to read it."* This is a security threat, as P11 sometimes works on clients' confidential information.

4.3.2 Employee's location could be traced

Some participants feared that some of the work calls made from home could be traced back to their location. This would result in the loss of their *privacy*. To illustrate, P9 works with the government and sometimes takes phone calls from angry citizens. While telecommuting, P9 uses her work mobile phone to make and receive calls from clients at home. P9 remarked: *"Sometimes, I wonder if [clients] are able to trace my phone calls. I know they're not [able to] because my [work] phone number doesn't pinpoint the exact location I am in. I work with [people's social insurance/security] numbers [and] addresses [on my system, and] a lot of the times when I get calls, some of them I realized have been close to my neighborhood. There was one call I received that was actually two streets down from where I was staying. And I [thought], '[What] if this person knew where I was located?' Sometimes I wonder, 'Oh, man, like if they knew where I was located, would they come to my house and ask me to do stuff?"* While this threat may be improbable, this fear made the participant anxious about handling work phone calls from home.

4.3.3 People might break into employee's house

There was a fear that someone could break into participants' houses to steal the company's equipment. This was a security concern and a constant fear for few participants who took home expensive work devices to aid telecommuting. If realized, this consequence of threat could result in the violation of participants' *privacy* and safety, loss of system *availability*, as well as data *confidentiality* and *integrity*, and, in extreme cases, the loss of *life*. For instance, P11 explained: *"My only other massive fear is, what if I had a break-in and somebody stole my [work] laptop? I mean, I have great confidence that that wouldn't happen, but it absolutely has been a fear. I think that's probably [the] only sort of ... situation that genuinely creates the occasional bit of anxiety for me ... 'Jesus, how do I know I am [secure]?' [Someone breaking in] seems like one of those improbable situations, but not impossible. So, even saying it out loud makes me nervous that somehow I am creating that reality now, because we certainly have people [in my neighborhood] with addictions who sooner or later need to feed their addictions and need to get money and sometimes get desperate."*

4.4 Human Dimension

These are challenges that were specific to individuals and their varying capabilities or limitations. We explain these challenges below.

4.4.1 Challenges with using the technology

Some participants were not tech-savvy, which made it harder for them to switch to full-time telecommuting. P7 (network engineer), for instance, remarked: *"The human aspect of security is always the biggest problem. [The IT personnel] are not there to monitor what everyone does at home on their computers all the time. Users don't know how to properly explain what their [technological] issue is; they use end-user terminology instead of technical terminology. So trying to translate the communication with the users was the biggest challenge. [When users had a technical issue,] trying to get them to explain to us what the problem [was challenging]."*

Lack of technical knowledge could lead to dangerous errors. This outcome of threat was particularly a concern when there was a disconnect between the participants' knowledge and what the organization expected them to do. For instance, some participants could fail to install security-critical software updates on their work systems while telecommuting, due to the lack of the technical capacity to do so. This challenge could lead to the loss of *integrity* and *confidentiality* of the organizational data, should employees' computers become targets of cyber-attacks.

The lack of technological competence was also reflected in poor understanding of security. For example, when discussing virtual private networks (VPNs), P1 (digital communications specialist) remarked: *"VPN, is ... something that secures your laptop. I just know [VPN] makes everything safe. You can't get hacked. You can't [have] none of that [hacking]. Everything's secured."* In this particular case, P1 assumed that once she connected to her employer's network using a VPN, everything

on her laptop was secure.

4.4.2 The challenge of distinguishing real organizational emails from phishing ones

Participants had difficulty distinguishing between real organizational emails and phishing ones. Sometimes, employees had been so much sensitized about phishing emails that they would classify real organizational emails as phishing. P7 shared an illustrative story: *"[Prior to working from home, my organization had [a] service that would do hands-on training [and send] out test fake emails to [employees]. If anyone clicked on [one of these fake emails], they'd get a warning, that [said], 'By the way, this is not real; this is a phishing email.' Now, [while employees have been working from home], we were sending out updates regarding viruses and anti-viruses and then people were reporting [them] as [phishing emails], not realizing it was a legitimate board email. [People have become] too paranoid."*

It was hard for some participants to recognize legitimate work-from-home precautions and apply them as needed. Some of these precautions are required to protect the confidentiality and integrity of work data. Therefore, similar to the challenges of using technology (§4.4.1), this challenge could lead to the loss of *confidentiality* and *integrity* of organizational data.

4.5 Organizational Dimension

The major challenge was that organizations sometimes provided few or no guidelines on how to telecommute. We define telecommuting guidelines as a set of instructions for employees about what to take home from work, how to set up their home office, and how to ensure the security and privacy of work-related information. We discuss this challenge below and explain how it led to other security and privacy issues for participants.

Many participants received little or no guidance on telecommuting. P15, for instance, was handling financial information while working from home. However, it was unclear to him how he would do that safely. When asked about guidelines regarding working from home, he explained: *"We barely get told anything [regarding telecommuting]. ... There hasn't been any communication with regard to how to handle confidential conversations over the phone. We just use our discretion [in handling financial] matters [over the phone].*

Telecommuting violates the organizations' work policies. For some organizations, working from home violates the organization's policies, and therefore, there are no guidelines for employees. When P11 was asked about the work-from-home guidelines instituted by his organization, he remarked: *"There were no guidelines [for telecommuting;] in fact, ... [working from home] is breaking [the] guidelines. ... We had just recently completed a very thick policy manual about data protection, information, privacy, [and] security ... that indicated [that] you don't take anything [from] work [to] home. All work will be done from the office. So in fact, having to respond to the pandemic created a conflict with recent policies around the security of information."* As such, people in some organizations had no guidelines on how to work from home.

Participants, therefore, came up with their own norms of working from home. They used their own understanding and interpretation of security and privacy best practices. For instance, when asked about her work-from-home practices, P3 explained: *"[Be]cause I'm working on my personal computer, [I'm] not saving anything on my actual computer a whole lot. ... I save everything on my [USB] stick. It's not too hard [to remember to save files on my USB stick] because I just leave the stick plugged into my computer ... so it's right there."* P3 further explained that she secured her laptop by using a password, though her USB stick was not encrypted or password protected. Since P3's USB stick was always plugged into the computer, the information saved on the USB stick was only as secure as the information saved on her personal computer or even less. The concern is that attackers (who could be household members) need a password to access the files saved on P3's laptop, but attackers can easily access the USB stick files. This challenge could lead to the organizational outcome of threat of the loss of data *confidentiality* if attackers had access to the USB stick.

5 Discussion

Our findings point to the security and privacy challenges, threats, and potential outcomes of threats that participants perceive while telecommuting. Figure 1 illustrates the consequences of a threat that could arise due to the identified challenges and threats with telecommuting, which we described in the previous section. In this section, we generalize discussion of the results in the form of perceived outcomes of threats to telecommuters and their employers. In Table 1 and Table 2, we present the challenges and threats, as perceived by participants, and show how they could lead to various outcomes of threats. We identified participants' perceived outcome of threat in which the organization's assets are at stake (Table 2). In contrast with office work, mass telecommuting introduces additional consequences of threats. The participants' privacy, data, and in some cases, well-being are at stake (Table 1). In the rest of the discussion section, we describe both types of these outcomes of threats and discuss options for mitigating some of them. It should be noted, however, that proper evaluation of these countermeasures is subject to future research.

While some of the challenges and threats are not unique to telecommuting, the issues are amplified in scale and severity when workers solely rely on telecommuting. The severity of the challenge gets intensified due to the lack of physical proximity among the coworkers for many weeks, if not months.

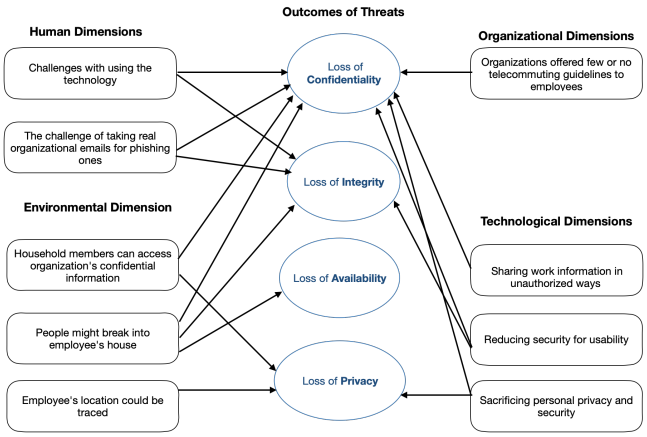


Figure 1: The relationship between challenges, threats, and the outcomes of threats. Arrows link challenges\threats to the outcomes of the threat.

For example, confidential information may have never been shared through unauthorized means (§4.2.1), because employees would meet in person. However, mass telecommuting takes away that opportunity, leaving employees with nothing else but to rely solely on online solutions, some of which (in isolation or in combination with other technologies) turned out to be over-restrictive or otherwise have less than acceptable usability (§4.2.3). Another example is the possibility of using technological tools to monitor employees' activities, which could result in an invasion of their privacy (§4.2.2). This challenge could lead to bigger issues, such as monitoring employees' or coworkers' daily routine even during weekends. These privacy issues became much more of a concern when long-term mass telecommuting became widespread overnight and might even remain so after the pandemic [15, 16, 32, 55]. Identifying and addressing these challenges, therefore, would go a long way toward improving telecommuting beyond the pandemic. Further, mass telecommuting could also happen in emergency situations such as power outages, earthquakes, and other natural and human-made disasters. In the rest of this section, we categorize recommendations into three types, according to the intended audience: organizations (R-O), employees (R-E), and those working with telecommuters (R-T).

5.1 Perceived Outcome of Threat Toward Workers

Telecommuting elevates the outcomes of threats to personal safety for employees and their households. Some participants worried that angry clients could locate their homes and terrorize them (§4.3.2). Other participants were anxious that criminals could break into their homes and steal their organizations' expensive work devices while also putting participants' privacy and safety at risk (§4.3.3). These anxieties could negatively affect employees' productivity, job satisfaction or employee retention while telecommuting [22, 38, 50]. Physical security at work is the responsibility of the employers and is commonly implemented by monitoring and controlling access to the office space and parking lots, and by stationing security personnel in the office buildings [10, 25, 63]. In the telecommuting scenario, however, the expensive work equipment now resides in the employees' homes, and there is no physical security provided. Organizations could implement encryption of the computer's hard drive to safeguard their data [29, 36, 61, 73]. However, the safety of the employees and the household members is also at risk due to telecommuting. Therefore, telecommuting produces a negative externality [30], as it is the employer that benefits from the employee being able to telecommute, but it is the household members who have to mitigate the elevated risk to physical safety and the psychological trauma that comes with it.

Employers can put measures in place to manage the safety of the telecommuters and their households **(R-O)**. Organizations need to be sensitive to the employees' physical security and consider the reality that different employees live in neighborhoods with varying safety levels (§4.3.3). Organizations can be mindful of this threat and manage it as part of their policies or processes for handling work from home. For long-term (and full-time) telecommuting, the employers could consider setting up home alarm systems for their employees. The employers could also look into setting up work hubs where the organization's devices could be set up and the employee's safety is protected. Further, employers can educate employees about security measures at the work hub to allay their fears. We also suggest that organizations provide clear guidelines on managing the home-work environment to optimize employees' physical safety. For instance, similar to on-site organizational security measures, employers could develop processes for physical security while telecommuting, such as help lines or safety routines that employees could use if the organization's clients/customers misuse employees' personal data.

Loss of workers' privacy is the major theme that emerged in the interviews. As can be seen in the rightmost column of Table 1, every type of concern is related to this theme. The main reason for its omnipresence, we believe, is that telecommuting is a hybrid work situation, where employees are at home but expected to carry out the organization's activities. Therefore, employees must behave in a specific way, which comes at the cost of their privacy. For instance, employees gave clients and coworkers (and even sometimes customers) their own phone numbers and other personal information (§4.2.2). The participants had other privacy boundaries (e.g., by answering phone calls from unknown numbers) compromised to facilitate telecommuting (§4.2.2). Workers were also worried about others taking screenshots of them without their consent during video calls (§4.2.2) and others feared that their clients and colleagues could overhear personal conversations taking

Table 1: Perceived Outcome of Threat Toward Workers

Asset	Employee's behavior	Threat agents	Reason for concern	Threat	Outcome of threat
1. Employee's personal phone number and home address	Employee giving coworkers their personal information to aid telecommuting	Coworkers	a. Violation of personal boundaries b. Less control over who has access to personal information	a. Coworkers could use employee's personal information for purposes other than initially declared b. Sharing of personal information without permission from the subject of the information	**a. Misuse and unauthorized sharing of shared personal data** **b. Loss of privacy** (§4.2.2)
2a. Employee's money b. Employee's privacy	Employee picking up calls from unknown numbers, not screening phone calls	Phone scammers	Reduced protection from scam calls	a. Phone scammers could obtain employee's financial information b. Increase in scam calls	**a. Abuse of personal data** **b. Becoming a victim of scams** **c. Loss of privacy** (§4.2.2)
3a. Employee's private home setting b. Housemates' privacy c. Employee's privacy	Employee forced to turn on their video camera during telecommuting	Coworkers	a. Personal environment of the employee is exposed to coworkers b. Lack of privacy in the home environment VS the work environment	a. Coworkers seeing employee's private environment and housemates b. Employee's improper disclosure of themselves	**a. Accidental disclosure** **b. Loss of privacy** (§4.2.2)
4. Employee's routine	Using technological tools that make it easy to monitor employees	Coworkers, managers	Coworkers and managers can monitor employee's activities and routine	Coworkers and managers could use this information to predict employee's routine	**Loss of privacy** (§4.2.2)
5. Employee's personal data	Giving students remote access to the employee's computer	Students	Due to a lack of computer knowledge, there is uncertainty about what students can do on the employee's laptop when given remote access via videoconferencing	Students could control the computer of a non–tech-savvy employee and access personal data	**a. Abuse of personal data** **b. Loss of privacy** (§4.2.2)
6. Employee's safety	Calling customer/client from home	Customer/client	Unmasked work phone number	An angry customer/client could locate employee's home by tracing phone calls made to the customer/client	**a. Abuse of personal data** **b. Loss of life** **c. Loss of privacy** (§4.3.2)
7. Employee's safety	Distributing care packages from home	Criminals present in neighborhood	Physical harm by intruders during a break-in to the house	Physical harm and injury	**a. Loss of life** **b. Loss of privacy** §4.3.3)

place at the workers' homes (§4.3.1).

There are various ways for employers to aid their employees in maintaining privacy while working from home. Organizations can provide some form of phone number masking (which prevents others from knowing the actual phone number of the caller) or VoIP solutions [42] to employees who have to use their personal phones for work [23] **(R-O)**. Further, we suggest technology support for alerting participants of video calls when screenshots are taken, to help employees maintain awareness of their privacy violations and to deter abuse of such capabilities by others (§4.2.2) **(R-E)**. To prevent clients and colleagues from hearing personal conversations happening in the household, teleconferencing software and phones could have a feature where the microphone is automatically muted when employees are not talking. Using voice recognition, the microphone automatically unmutes when the employee starts talking to the client or coworker **(R-T)**. There could also be directional microphones on phones and videoconferencing apps, whereby the technology only picks up the voice of the person in front of the computer or phone **(R-T)**.

Furthermore, there seems to be a conflict between employees maintaining their privacy and doing their job. Our findings confirm Pyöriä's work, as this author predicted disruption to privacy in employees' homes as a challenge that could arise in teleworking [52]. Our participants experienced a dilemma around whether to turn on their webcams during work meetings. For some, turning on the webcams was an invasion of privacy, as it welcomed coworkers into their private homes and lives. On the other hand, employers expected participants to always have their webcams on during work meetings as these meetings are done within work hours (§4.2.2). Further, some employees also had to give clients remote access to their personal computers while telecommuting (§4.2.2). In addition, some telecommuting solutions could aid with monitoring employees' activities and detect when employees were at or away from their desks (§4.2.2). Research shows that such online status indicators or presence sharing applications leads to privacy concerns for users [12,18,28,56]. Other features of videoconferencing apps raise further concerns about employees' privacy during telecommuting. For instance, Microsoft Teams and Zoom allows meeting participants to livestream a meeting without getting consent from the participants [69,81]. Therefore, employees' work meetings in their personal spaces can be livestreamed on Facebook Live and YouTube without the employees' knowledge. All of these situations raise questions about employers' rights over employees privacy in their own homes. Palen et al. discussed the issues surrounding privacy in a technologically connected world. Because privacy is personal, people set various boundaries in their everyday life to maintain their privacy [3,49,71]. However, the use of information technology disrupts or demolishes those boundaries. The authors explain the challenge further: "problems emerge when participation in the networked world is not deliberate, or when the bounds of identity definition are not within one's total control. [49]" As seen in our results, employees do not have full control over their privacy, which is a challenge. There is also the issue of content collapse in telecommuting. "The concept of context collapse describes the process by which connections from various aspects of individuals' lives become grouped together under generic terms [1,11,74]." Similarly, in

telecommuting, workers experience context collapse and are faced with the dilemma of how to draw boundaries between their personal and professional lives. This leads to privacy issues for participants (§4.2.2).

To help create a balance between privacy and doing one's job, organizations can have discussions and transparency on how much privacy employees are entitled to when telecommuting **(R-O)**. It may be helpful for organizations to clearly state what they expect from employees regarding having the camera on or off while working from home, dress code while telecommuting, or giving clients their personal phone numbers. There might, however, be no clear-cut answers to these questions. Moreover, they raise bigger questions that future research could look into. For example, can employees maintain their privacy while working from home? If yes, how can privacy boundaries be maintained while respecting organizational cultures, social norms, and work policies? Does the use of technologies that monitor employees' routines (mostly during work hours) violate their privacy? Should technological tools be allowed to monitor workers' activities during and after work hours when they work from home? How can employees give or withdraw their consent for recording, screenshots, or livestreaming during online work meetings without feeling stigmatized or fearing repercussions? How can organizations and technologists make sure employees are not putting their physical safety at risk when working from home (§4.2.2)? Employers and employees need to consider these different scenarios when making telecommuting arrangements.

5.2 Perceived Outcome of Threat Toward Organizations

The outcomes of threats related to the confidentiality and integrity of the organization's assets were the most common theme in this category (see Table 2). Kintner et al. and Spinellis et al.'s participants also predicted inadequate security for protecting transmitted information in teleworking as a potential challenge [34, 66]. In some cases, the organization's assets were at risk because the official work communication platforms' were not usable (§4.2.3). Therefore, participants used other insecure but usable and familiar technological solutions to talk to coworkers and share clients' confidential information. Since participants no longer had the luxury of talking in person to their colleagues about work-related matters, participants were looking for technology support closest to in-person interactions. Such support made communication with coworkers easy without unnecessary setup or complicated authentication procedures (§4.2.1 & §4.2.3). Employers need to ensure that work communication platforms are very intuitive and easy, if they want to address this issue **(R-O)**. These work communication platforms could also be linked to other popular social communication channels. For example, organizations could work toward having a secured platform on Facebook to discuss work-related information. One of the principles of secure systems design is the path of least resistance [79]. This principle states that "to the greatest extent possible, the natural way to do any task should also be the secure way" [79]. Since employees are already using these social platforms anyway, employees are most likely to follow the path of least resistance. Such types of platforms are subject to future research and development.

The inability to distinguish between phishing and real emails rendered employers' announcements ineffective. Some organizations asked their employees to use their personal devices to work and expected employees to use the organization's software on those devices. Because IT personnel didn't have control or access to the employees' devices, IT personnel had to send emails to the employees with system updates required to maintain the organization's software while telecommuting. Because employees found it challenging to distinguish between fake and real emails, employees ignored important system updates sent through emails (§4.4.2). Organizations could make use of already existing solutions to digitally sign and encrypt official emails from the organizations [47] **(R-O)**. Employees would, however, need to learn and understand how these solutions work because, as previous research shows, people find it difficult to use encrypted and signed emails correctly [76]. Apart from email, we suggest that other communication platforms could be used, such as a usable official messaging platform to relate work information **(R-O)**.

There was also the outcome of threat of household members overhearing confidential work discussions. In real-life situations, these confidential conversations are mostly held in offices, which are considered safe enough for those conversations to happen. However, in the context of telecommuting, home environments do not necessarily provide sufficient sound insulation. While this might not be an acute issue for traditional households with one family, cohousing [13], collective housing, and similar arrangements that are increasingly common in urban areas where housing is expensive significantly decrease control and awareness of who might be in a household and possibly overhear discussions at any given moment.

There is no easy way to address this problem. The solution is not as simple as telling employees to take work calls where other household members cannot overhear the conversation. By default, there seems to be an assumption that the employee's home environment is a typical family setting with father, mother, and child(ren) and an office space with a closed door where the employee can conveniently take work calls. In reality, employees have a wide variety of cohabitation arrangements and environments and for some, it is simply impossible to avoid working in a space shared with the housemates. Further, in some cases working in a separate room doesn't solve the problem of poor sound insulation (§4.3.1). Organizations **(R-O)** need to be sensitive to the fact that employees' living situations vary and should be mindful of the corresponding

Table 2: Perceived Outcome of Threat Toward Organizations

Asset	Employee's behavior	Threat agents	Reason for concern	Threat	Outcome of threat
1. Confidential information	Putting organizations' and customers' confidential information on social media platforms	Employees of social media platforms, cybercriminals	Lack of confidentiality on social media platforms	a. Employees of the social media platform could spy on the organization's confidential data b. Cybercriminals could exploit the vulnerabilities of social media platforms and obtain confidential information	**Loss of confidentiality** (§4.2.1)
2. Customer/client's confidential information	a. Discussing confidential information through device speakers b. Reading out clients' confidential information	Housemates	Lack of sound insulation	Housemates could overhear confidential information	**Loss of confidentiality** (§4.2.1)
3a. Citizen's information b. Political report that has not been made public	Making use of a less secure personal phone and email software	Social insiders, cybercriminals	Personal phones and email software are not configured to be as secure as work phones and emails	a. Social insiders snooping through employee's phone and accessing their text messages b. Hijacking personal email account and obtaining copies of the work emails	**Loss of confidentiality** (§4.2.3)
4. Organization's accounting information	Reducing the security of systems to aid telecommuting	Cybercriminals	Reduced security of remote desktop server	Cybercriminals could compromise the security of the system and access organization's data	**a. Loss of confidentiality** **b. Loss of integrity** (§4.2.3)
5. Confidential information	Giving students remote access to employee's personal computer	Students	Due to a lack of computer knowledge, there is uncertainty about what students can do on the employee's laptop when given remote access via videoconferencing	Student could control the computer of a no–tech-savvy employee and access confidential data	**a. Loss of confidentiality** **b. Loss of integrity** §4.2.2)
6. Client's health information	Displaying confidential information on big screens, in large font sizes, while telecommuting in the kitchen area	Housemates	Housemates could read confidential information off the screen	Housemates could view confidential health information	**Loss of confidentiality** (§4.3.1)
7. Organization's confidential information	Unable to troubleshoot work devices from home	Cybercriminals	Reduced security of work devices for telecommuting	Cybercriminals could exploit vulnerabilities in work devices	**Loss of confidentiality** (§4.4.1)
8. Organization's confidential information	Using expensive organizational work devices to aid telecommuting	Criminals present in neighborhood	Lack of physical security of work devices and recent break-in	Neighbors could break into employee's home and steal work equipment	**a. Loss of confidentiality** **b. Loss of integrity** **c. Loss of availability** (§4.3.3)

outcomes of threats to the confidentiality of work calls.

There is a need for discourse in the research community on the possible solutions to these problems. Table 1 and Table 2 present a comprehensive illustration of possible outcomes of threats to organizations and employees while telecommuting. As telecommuting becomes more full-time and long-term [15, 16, 24, 32, 41, 55], the topics and issues surrounding organizational data security and employees' safety and privacy need to be discussed and addressed. The main topic is that there is a dilemma around employees maintaining their privacy and safety while telecommuting and employers ensuring that employees carry out their work from home and safeguard their organization's data. With the increase in successful cyber-attacks on telecommuters [6, 33, 40, 77], addressing the identified security and privacy challenges and threats encountered by employees may go a long way in reducing cyber-attacks related to telecommuting. We believe our study provides insights into these challenges and serves as a basis for possible solutions to be explored and discussed and will ultimately lead to better work-from-home practices for both employees and employers.

5.3 Limitations

Our sample could have been more balanced and diverse. It had more male (56%) participants, though statistics show that more men are employed than women [48]. The average and median age were 41 and 38, respectively. We could have recruited more older participants. However, the oldest participant was 64, and statistics show that on average the age of retirement is 62 [14, 62]. Further, we do not have enough data about the context (e.g., participant's environment) to determine whether each threat is realistic or probable. As with most qualitative research, the data were self-reported and may have been affected by several systematic biases such as social desirability, halo effect, and acquiescence response bias [20]. Nonetheless, we believe that our study results can serve as a background for further research and discourse on how to improve the security and privacy of telecommuters.

6 Acknowledgements

This research has been supported in part by a gift from Scotiabank to UBC. We thank Larisa Lensink for having brainstorming sessions with the lead author on the idea of doing this study. We appreciate all participants involved in the study. We thank members of the University of British Columbia's Laboratory for Education and Research in Secure Systems Engineering, which provided feedback on the reported research and the earlier versions of the paper. We thank our anonymous reviewers for their feedback and suggestions for improving the paper. Stylistic and copy editing by Eva van Emden helped to enhance the readability of this paper.

References

[1] Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. Privacy and human behavior in the age of information. *Science*, 347(6221):509–514, 2015.

[2] Abi Adams-Prassl, Teodora Boneva, Marta Golin, and Christopher Rauh. Work tasks that can be done from home: Evidence on variation within & across occupations and industries. *CEPR Discussion Paper No. DP14901*, 2020.

[3] Irwin Altman. The environment and social behavior: privacy, personal space, territory, and crowding. *ERIC*, 1975.

[4] John Ameriks, Joseph Briggs, Andrew Caplin, Minjoon Lee, Matthew D Shapiro, and Christopher Tonetti. Older americans would work longer if jobs were flexible. *American Economic Journal: Macroeconomics*, 12(1):174–209, 2020.

[5] Shelly Banjo, Livia Yap, Colum Murphy, and Vinicy Chan. Coronavirus forces world's largest work-from-home experiment. https://www.bloomberg.com/news/articles/2020-02-02/coronavirus-forces-world-s-largest-work-from-home-experiment, 2020. Accessed: 2020-09-11.

[6] Carbon Black. Global threat report extended enterprise under threat. https://www.carbonblack.com/wp-content/uploads/VMWCB-Report-GTR-Extended-Enterprise-Under-Threat-Global.pdf, 2020. Accessed: 2020-09-11.

[7] Nicholas Bloom. The bright future of working from home. https://siepr.stanford.edu/research/publications/bright-future-working-home, 2020. Accessed: 2020-09-11.

[8] Nicholas Bloom. How working from home works out. https://siepr.stanford.edu/research/publications/how-working-home-works-out, 2020. Accessed: 2020-09-11.

[9] Nicholas Bloom, James Liang, John Roberts, and Zhichun Jenny Ying. Does working from home work? evidence from a chinese experiment. *The Quarterly Journal of Economics*, 130(1):165–218, 2015.

[10] Security Boulevard. Best practices: 6 physical security measures every company needs. https://securityboulevard.com/2019/03/best-practices-6-physical-security-measures-every-company-needs/, 2019. Accessed: 2021-02-1.

[11] Danah Boyd. Taken out of context: American teen sociality in networked publics. *Available at SSRN 1344756*, 2008.

[12] Andreas Buchenscheit, Bastian Könings, Andreas Neubert, Florian Schaub, Matthias Schneider, and Frank Kargl. Privacy implications of presence sharing in mobile messaging applications. In *Proceedings of the 13th international conference on mobile and ubiquitous multimedia*, pages 20–29, 2014.

[13] Cohousing Califonia. Cohousing califonia. https://www.calcoho.org, 2021. Accessed: 2021-02-24.

[14] Statistics Canada. Retirement age by class of worker. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410006001, 2020. Accessed: 2020-09-11.

[15] CBC. Shopify permanently moves to work-from-home model. https://www.cbc.ca/news/canada/ottawa/shopify-pandemic-staff-ottawa-1.5578614, 2020. Accessed: 2021-01-18.

[16] Katie Clarey. In the next decade, half of facebook's workforce could be remote. https://www.hrdive.com/news/facebook-remote-workforce-zuckerberg-announcement/578578/, 2020. Accessed: 2021-01-18.

[17] Fintan Clear and Keith Dickson. Teleworking practice in small and medium-sized firms: management style and worker autonomy. *New Technology, Work and Employment*, 20(3):218–233, 2005.

[18] Camille Cobb, Lucy Simko, Tadayoshi Kohno, and Alexis Hiniker. User experiences with online status indicators. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

[19] Deborah Cohen and Benjamin Crabtree. Qualitative research guidelines project. http://www.qualres.org/, 2006.

[20] Diane Dodd-McCue and Alexander Tartaglia. Self-report response bias: Learning how to live with its diagnosis in chaplaincy research. *Chaplaincy Today*, 26(1):2–8, 2010.

[21] Gus Evangelakos. Keeping critical assets safe when teleworking is the new norm. *Network Security*, 2020(6):11–14, 2020.

[22] Michael T Ford, Christopher P Cerasoli, Jennifer A Higgins, and Andrew L Decesare. Relationships between psychological, physical, and behavioural health and work performance: A review and meta-analysis. *Work & Stress*, 25(3):185–204, 2011.

[23] Derek Frome. Masked calling. https://www.twilio.com/docs/glossary/what-is-masked-calling, 2020. Accessed: 2020-01-07.

[24] Ryan Golden. Gartner: Over 80% of company leaders plan to permit remote work after pandemic. https://www.hrdive.com/news/gartner-over-80-of-company-leaders-plan-to-permit-remote-work-after-pande/581744/, 2020. Accessed: 2021-01-18.

[25] Greetly. Workplace security & access control - the fundamentals. https://www.greetly.com/blog/workplace-security-access-control-the-fundamentals, 2020. Accessed: 2021-02-1.

[26] Gregory Guest, Kathleen M MacQueen, and Emily E Namey. *Applied thematic analysis*. Sage Publications, 2011.

[27] Gregory Guest, Kathleen M MacQueen, and Emily E Namey. Introduction to applied thematic analysis. *Applied thematic analysis*, 3:20, 2012.

[28] Roberto Hoyle, Srijita Das, Apu Kapadia, Adam J Lee, and Kami Vaniea. Was my message read? privacy and signaling on facebook messenger. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 3838–3842, 2017.

[29] Huridocs. 5 steps to protect your data in case of computer theft. https://huridocs.org/2015/12/steps-to-protect-your-data-computer-theft/, 2015. Accessed: 2021-02-1.

[30] Corporate Finance Institute. What are negative externalities? https://corporatefinanceinstitute.com/resources/knowledge/economics/negative-externalities/, 2021. Accessed: 2021-02-24.

[31] INTERPOL. Interpol report shows alarming rate of cyberattacks during covid-19. https://www.interpol.int/en/News-and-Events/News/2020/INTERPOL-report-shows-alarming-rate-of-cyberattacks-during-COVID-19, 2020. Accessed: 2021-01-18.

[32] Jack Kelly. Twitter ceo jack dorsey tells employees they can work from home 'forever'—before you celebrate, there's a catch. https://www.forbes.com/sites/jackkelly/2020/05/13/twitter-ceo-jack-dorsey-tells-employees-they-can-work-from-home-forever-before-you-celebrate-theres-a-catch/?sh=771246c82e91, 2020. Accessed: 2021-01-18.

[33] Yiftach Keshet. Recent escalations in cyberattacks in italy prove the coronavirus impact on cybersecurity – acting as a warning for cisos worldwide. https://www.cynet.com/blog/recent-escalation-in-cyberattacks-in-italy-prove-the-coronavirus-impact-on-cybersecurity-acting-as-a-warning-for-cisos-worldwide/, 2020. Accessed: 2020-09-11.

[34] Susan Kintner. Preliminary report telework/telecommuting: Employers' perspectives and perspectives of service members and veterans with disabilities. *e-Networks in an Increasingly Volatile World*, page 204, 2006.

[35] D Richard Kuhn, Miles C Tracy, and Sheila E Frankel. Security for telecommuting and broadband communications. *NIST Special Publication*, 800:46, 2002.

[36] Micah Lee. Encrypting your laptop like you mean it. https://theintercept.com/2015/04/27/encrypting-laptop-like-mean/, 2021. Accessed: 2021-02-1.

[37] Dan Lohrmann. 2020: The year the covid-19 crisis brought a cyber pandemic. https://www.govtech.com/blogs/lohrmann-on-cybersecurity/2020-the-year-the-covid-19-crisis-brought-a-cyber-pandemic.html, 2020. Accessed: 2021-01-18.

[38] Julie M McCarthy, John P Trougakos, and Bonnie Hayden Cheng. Are anxious workers less productive workers? it depends on the quality of social exchange. *Journal of Applied Psychology*, 101(2):279, 2016.

[39] Google Meet. Change your background in google meet. https://support.google.com/meet/answer/10058482?co=GENIE.Platform%3DDesktop&hl=en&oco=1#zippy=%2Cwhy-dont-i-have-the-change-background-option, 2021. Accessed: 2021-02-19.

[40] Ezequiel Minaya. 4,000% increase in ransomware emails during covid-19. https://www.nationalobserver.com/2020/04/14/news/4000-increase-ransomware-emails-during-covid-19?utm_source=National+Observer&utm_campaign=71c5787b54-EMAIL_CAMPAIGN_2020_04_14_12_21&utm_medium=email&utm_term=0_cacd0f141f-71c5787b54-276991505, 2020. Accessed: 2020-09-11.

[41] Ezequiel Minaya. Cfos plan to permanently shift significant numbers of employees to work remotely — survey. https://www.forbes.com/sites/ezequielminaya/2020/04/03/cfos-plan-to-permanently-shift-significant-numbers-of-employees-to-work-remotely---survey/#11bc806575b2, 2020. Accessed: 2020-09-11.

[42] Nextiva. Protect your number with call masking. https://www.nextiva.com/features/voip/call-masking.html, 2021. Accessed: 2021-02-24.

[43] Marian Niedźwiedziński and Anna Bąkała. Telework and security. *Systems: journal of transdisciplinary systems science*, 12(1), 2007.

[44] Insurance Bureau of Canada. Cyber risks: An increased threat during covid-19. http://www.ibc.ca/on/business/risk-management/cyber-risk/an-increased-threat-during-covid-19, 2020. Accessed: 2021-01-18.

[45] U.S. Bureau of Labour Statistics. Economic news release. https://www.bls.gov/news.release/flex2.htm, 2020. Accessed: 2020-09-11.

[46] Kenneth Okereafor and Phil Manny. Solving cybersecurity challenges of telecommuting and video conferencing applications in the covid-19 pandemic. *Journal Homepage: http://ijmr. net. in*, 8(6), 2020.

[47] OpenPGP. Openpgp about. https://www.openpgp.org, 2021. Accessed: 2021-02-1.

[48] International Labour Organisation. The gender gap in employment: What's holding women back? https://www.ilo.org/infostories/en-GB/Stories/Employment/barriers-women#intro, 2018. Accessed: 2020-09-11.

[49] Leysia Palen and Paul Dourish. Unpacking" privacy" for a networked world. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 129–136, 2003.

[50] I Plaisier, ATF Beekman, R De Graaf, JH Smit, R Van Dyck, and BWJH Penninx. Work functioning in persons with depressive and anxiety disorders: the role of specific psychopathological characteristics. *Journal of affective disorders*, 125(1-3):198–206, 2010.

[51] Choudhury Prithwiraj, Cirrus Foroughi, and Barbara Larson. Work-from-anywhere: The productivity effects of geographic flexibility. *Strategic Management Journal*, 42(4):655–683, 2021.

[52] Pasi Pyöriä. Knowledge work in distributed environments: issues and illusions. *New Technology, Work and Employment*, 18(3):166–180, 2003.

[53] Pasi Pyöriä. Managing telework: risks, fears and rules. *Management Research Review*, 2011.

[54] André Queirós, Daniel Faria, and Fernando Almeida. Strengths and limitations of qualitative and quantitative research methods. *European Journal of Education Studies*, 2017.

[55] Miguel Quiroga. Visible's shift to a permanent work from home model. https://www.linkedin.com/pulse/visibles-shift-permanent-work-from-home-model-miguel-quiroga/?trackingId=28Ku%2F%2Ft4b5g4PVyiJ9PBsA%3D%3D, 2020. Accessed: 2021-01-18.

[56] Yasmeen Rashidi, Kami Vaniea, and L Jean Camp. Understanding saudis' privacy concerns when using whatsapp. In *Proceedings of the Workshop on Usable Security (USEC'16)*, pages 1–8, 2016.

[57] Reddit. Black screen or entire screen flickering. https://www.reddit.com/r/Zoom/comments/fyi2ip/black_screen_or_entire_screen_flickering/, 2020. Accessed: 2020-09-17.

[58] Brie Weiler Reynolds. Differences between teleworking and telecommuting. https://www.flexjobs.com/blog/post/telecommuting-or-telework-whats-the-difference/, 2020. Accessed: 2020-09-17.

[59] Ron Ross, Michael McEvilley, and Janet Oren. Systems security engineering: Considerations for a multidisciplinary approach in the engineering of trustworthy secure systems. Technical report, National Institute of Standards and Technology, 2016.

[60] Katrin Schmelz and Anthony Ziegelmeyer. Reactions to (the absence of) control and workplace arrangements: experimental evidence from the internet and the laboratory. *Experimental economics*, pages 1–28, 2020.

[61] Jack Schofield. How can I protect my data if my laptop is stolen. https://www.theguardian.com/technology/2016/jul/07/how-can-i-protect-my-data-if-my-laptop-is-stolen, 2016. Accessed: 2021-02-1.

[62] USA Social Security. Retirement benefits. https://www.ssa.gov/benefits/retirement/planner/agereduction.html, 2020. Accessed: 2020-09-11.

[63] Hashim Shaikh. The importance of physical security in the workplace. https://resources.infosecinstitute.com/topic/importance-physical-security-workplace/, 2018. Accessed: 2021-02-1.

[64] R. Shirey. Rfc 4949-internet security glossary, version 2. https://tools.ietf.org/html/rfc4949, 2007. Accessed: 2021-02-24.

[65] Skype. How do i customize my background for skype video calls? https://support.skype.com/en/faq/fa34896/how-do-i-customize-my-background-for-skype-video-calls, 2021. Accessed: 2021-02-19.

[66] Diomidis Spinellis, Spyros Kokolakis, and Stefanos Gritzalis. Security requirements, risks and recommendations for small enterprise and home-office environments. *Information Management & Computer Security*, 1999.

[67] Alice Sturgeon. Telework: threats, risks and solutions. *Information Management & Computer Security*, 1996.

[68] Microsoft Teams. Change your background for a teams meeting. https://support.microsoft.com/en-us/office/change-your-background-for-a-teams-meeting-f77a2381-443a-499d-825e-509a140f4780, 2021. Accessed: 2021-02-19.

[69] Microsoft Teams. How to live stream microsoft teams meeting to youtube, facebook live & others. https://www.youtube.com/watch?v=fGMYvHrIB6M, 2021. Accessed: 2021-02-1.

[70] Microsoft Teams. User presence in teams. https://docs.microsoft.com/en-us/microsoftteams/presence-admins, 2021. Accessed: 2021-02-1.

[71] Zeynep Tufekci. Can you see me now? audience and disclosure regulation in online social network sites. *Bulletin of Science, Technology & Society*, 28(1):20–36, 2008.

[72] Paul C van Oorschot. *Computer Security and the Internet*. Springer, 2020.

[73] VeraCrypt. Veracrypt about. https://www.veracrypt.fr/en/Home.html, 2021. Accessed: 2021-02-1.

[74] Jessica Vitak, Cliff Lampe, Rebecca Gray, and Nicole B Ellison. "why won't you be my facebook friend?" strategies for managing context collapse in the workplace. In *Proceedings of the 2012 iConference*, pages 555–557, 2012.

[75] Natalie Whaley. Surveillance in employment: The case of teleworking. *Technical Communication*, 47(2):260–260, 2000.

[76] Alma Whitten and J Doug Tygar. Why johnny can't encrypt: A usability evaluation of pgp 5.0. In *USENIX Security Symposium*, volume 348, pages 169–184, 1999.

[77] WHO. Who reports fivefold increase in cyber attacks, urges vigilance. https://www.who.int/news-room/detail/23-04-2020-who-reports-fivefold-increase-in-cyber-attacks-urges-vigilance, 2020. Accessed: 2020-09-11.

[78] May Wong. Stanford research provides a snapshot of a new working-from-home economy. https://news.stanford.edu/2020/06/29/snapshot-new-working-home-economy/, 2020. Accessed: 2020-09-11.

[79] Ka-Ping Yee. User interaction design for secure systems. In *International Conference on Information and Communications Security*, pages 278–290. Springer, 2002.

[80] Zoom. Virtual background. https://support.zoom.us/hc/en-us/articles/210707503-Virtual-background, 2020. Accessed: 2020-09-17.

[81] Zoom. Live streaming meetings or webinars using a custom service. https://support.zoom.us/hc/en-us/articles/115001777826-Live-Streaming-Meetings-or-Webinars-Using-a-Custom-Service, 2021. Accessed: 2021-02-1.

Appendices

A Participants' Demographics

ID	Age	Gender	Educational level	Place of work	Position at work	Number of employees	Location
P1	24	F	Bachelor's	University	Digital communications specialist	-	Montreal, Quebec
P2	32	F	Master's	Library	Manager of marketing and communications	-	Montreal, Quebec
P3	31	F	Master's	University	Research assistant	14	Kitchener, Ontario
P4	36	F	Master's	Community organization	Occupational therapist	2,000+	Mount Pearl, Newfoundland
P5	49	F	Master's	IT firm	Sales director	10,000+	Caledonia, Ontario
P6	51	M	Bachelor's	Provincial government	Senior staff	-	Halifax, Nova Scotia
P7	47	M	High school	High school	Network engineer	11,000	Mono, Ontario
P8	61	M	Bachelor's	Children's science museum	Accounting supervisor	101	Vancouver, British Columbia
P9	24	F	Bachelor's	Federal tax agency	Call center agent	40,000	Ottawa, Ontario
P10	38	M	Bachelor's	Realtor	Mortgage broker	11,000	Mono, Ontario
P11	52	M	Bachelor's	Community center	Health director	85	Port Hardy, British Columbia
P12	31	M	College	Telecommunications	Account manager	-	-
P13	25	F	Bachelor's	Car sharing service	Business operations manager	22,000	Vancouver, British Columbia
P14	64	M	Bachelor's	Cannabis producer	Call center representative	37,000+	New Maryland Parish, New Brunswick
P15	31	M	Bachelor's	Telecommunications company	Customer service rep	37,000+	New Maryland Parish, New Brunswick
P16	47	M	Master's	Public transport services	Director in the planning department	4,000	Toronto, Ontario
P17	38	F	Master's	Elementary school	Education assistant	-	Dawson Creek, British Columbia
P18	38	M	College	Arts and culture management organization	Executive director	3	Vancouver, British Columbia
P19	43	M	Bachelor's	University	Business support analyst	10,000+	Vancouver, British Columbia
P20	30	M	Bachelor's	College	Assistant registrar systems and reporting	-	-
P21	53	M	Master's	Securities commission	Senior project manager	-	-
P22	48	M	College	Telecommunications provider	Customer service call agent	-	-
P23	59	F	College	High school	School secretary	-	-
P24	24	F	College	High school	School teacher	35	Halifax, Nova Scotia

Table A.1: **Demographics of participants.**

B Summarized Recommendations to Organizations (R-O), Employees (R-E), and Those Working with Telecommuters (R-T)

Table B.1: Summarized recommendations.

Recommendations	R-O, R-E, R-T
1. Organizations could make use of already existing solutions to digitally sign and encrypt official emails from the organizations	R-O
2. Apart from email, we suggest that other communication platforms could be used, such as a usable official messaging platform to relate work information	R-O
3. Organizations need to be sensitive to the fact that employees live in various living conditions and mindful of the corresponding outcomes of threats to the confidentiality of work calls	R-O
4. Employers can put measures in place to manage the safety of the telecommuters and their households	R-O
5. Organizations can provide some form of phone number masking (which prevents others from knowing the actual phone number of the caller) or VoIP solutions to employees who have to use their personal phones for work	R-O
6. To help create a balance between privacy and doing one's job, organizations can have discussions and transparency on how much privacy employees are entitled to when telecommuting	R-O
7. Employers need to ensure that work communication platforms are very intuitive and easy, if they want to address this issue	R-O
8. Technology support for alerting participants of video calls when screenshots are taken, to help employees maintain awareness of their privacy violations and to deter abuse of such capabilities by others	R-E
9. To prevent clients and colleagues from hearing personal conversations happening in the household, teleconferencing software and phones could automatically mute the microphone when employees are not talking; using voice recognition, the microphone automatically unmutes when the employee starts talking	R-T
10. There could also be directional microphones on phones and videoconferencing apps, whereby the technology only picks up the voice of the person in front of the computer or phone	R-T

C Saturation Graph

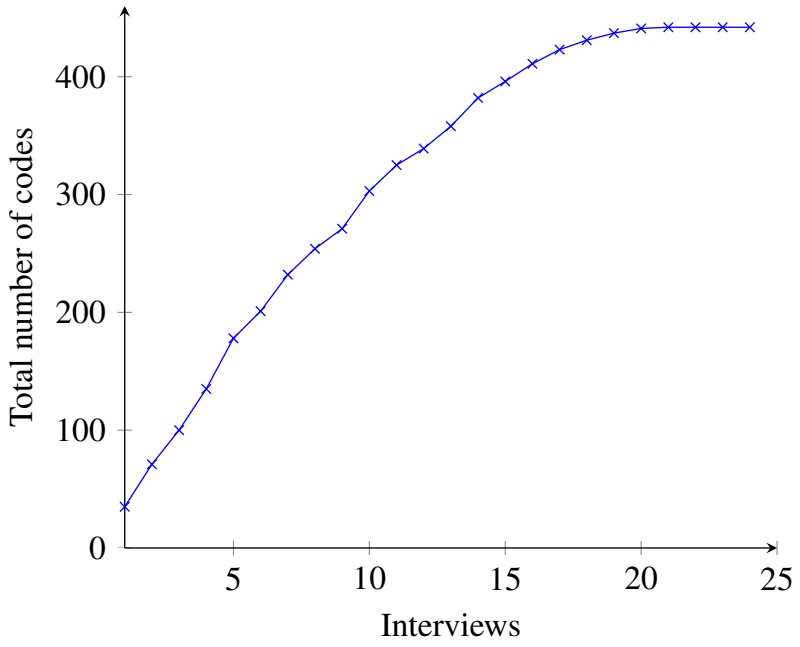


Figure C.1: Total number of codes after each interview.

D Interview Guide

Demographic questions

1. Age
2. Gender
3. Educational level
4. Place of work
5. Position at work

Interview questions

1. How has the pandemic affected your life?
2. How has it changed your life?
3. How has it changed your life in relation to others living with you?
4. Describe your typical work day before the pandemic
5. How many hours do you work?
6. Did you work remotely from home before the pandemic? If yes, how frequently?
7. How long have you been working from home?
8. If yes to the above question, how does your current remote work differ from previous experiences?

9. What is your experience working from home?
10. Describe your day-to-day work activities from home?
11. How does your current work activities differ from working in your physical office space?
12. What technology (or software, machines, devices) did you use to work with in the physical office space?
13. How do you handle/work with confidential communications in your work environment?
14. How do you manage confidential documents in your work environment?
15. What, if anything, is your workplace's guide on handling confidential communications and documents from home?
16. What, if anything, are the measures taken to comply with the organization's work-at-home rules?
17. What makes it easy to comply with these rules?
18. What makes it difficult to comply with these rules?
19. What motivates you, if anything, to be secured when you work from home?
20. What are your concerns, if any, with working from home as opposed to working in the office?
21. List the new technologies or software used specifically to work from home
22. What software or technologies have you explored since working from home?
23. If not mentioned ask, what video conferencing softwares have you been using to work from home?
24. If not mentioned ask, do you use VPNs to access your organization's resources?

For each technology used for remote working, ask:

25. Why did you choose this technology?
26. What if anything makes it easy to use the software? Why?
27. What if anything makes it difficult/complex to use the software?
28. If any complexity is discussed: How do you mitigate the complexities of using these technologies?
29. How does the technology assist you in securing your organization's confidential documents and communications?
30. How does the technology assist you in complying with your company's guide on protecting confidential documents and communications?
31. If you could change how the technology currently works, what would you change and why?
32. How do you handle concerns related to people in the household?
33. How is your current work environment?
34. What, if anything, would you like to change about your current work environment?
35. What other information do you think will be useful for this research?

Understanding Privacy Attitudes and Concerns Towards Remote Communications During the COVID-19 Pandemic

Pardis Emami-Naeini, Tiona Francisco, Tadayoshi Kohno, Franziska Roesner
Paul G. Allen School of Computer Science & Engineering
University of Washington

Abstract

Since December 2019, the COVID-19 pandemic has caused people around the world to exercise social distancing, which has led to an abrupt rise in the adoption of remote communications for working, socializing, and learning from home. As remote communications will outlast the pandemic, it is crucial to protect users' security and respect their privacy in this unprecedented setting, and that requires a thorough understanding of their behaviors, attitudes, and concerns toward various aspects of remote communications. To this end, we conducted an online study with 220 worldwide Prolific participants. We found that privacy and security are among the most frequently mentioned factors impacting participants' attitude and comfort level with conferencing tools and meeting locations. Open-ended responses revealed that most participants lacked autonomy when choosing conferencing tools or using microphone/webcam in their remote meetings, which in several cases contradicted their personal privacy and security preferences. Based on our findings, we distill several recommendations on how employers, educators, and tool developers can inform and empower users to make privacy-protective decisions when engaging in remote communications.

1 Introduction

The world was hit by a pandemic caused by the novel coronavirus and the COVID-19 disease in December 2019. In an attempt to prevent the spread of the virus, businesses and schools around the globe shut down, and people began sheltering in their homes to practice social or physical distancing [1].

Following social distancing protocols, people around the world have been encouraged or ordered to stay at home [2–4]. Hence, they started to work from home [5], keep in touch with family and friends remotely [6], and/or take remote courses [7], many for the first time [8]. This made remote conferencing tools an essential part of people's day-to-day lives, which albeit useful, posed potential privacy risks to people who are now regularly streaming video and audio from their own homes [9–15].

Increasingly integrated into people's lives and routines, widespread remote communications will not disappear with the end of the pandemic [16]. As people continue to work, socialize, and learn from home, it becomes imperative for their privacy and security to be protected. This requires the designers of in-home technologies (e.g., conferencing tools) and organizations that use them to understand the diverse needs of users. Understanding users' needs will enable designers and organizations to i) inform users about potential risks, and ii) gear their designs toward enabling users to control their privacy and security when using such technologies.

To that end, we conducted a worldwide survey (n =220) on Prolific [17] in May 2020, i.e., a few months into the pandemic, as people were newly settling into widespread remote communications. We sought to conduct our study during the transition phase of the COVID-19 pandemic, when participants were still adjusting to their new remote settings, while remembering their normal lives before the pandemic. Our survey covered three contexts of remote communications, namely work from home (WFH), socialize from home (SFH), and learn from home (LFH). In each context, without priming participants by asking directly about privacy and security, we leveraged participants' open-ended responses to tease out their unbiased privacy and security attitudes and behaviors towards three aspects of remote communications: conferencing tools, modes of remote communications (microphone, webcam), and locations of remote communications. We conducted quantitative and qualitative analyses to answer the following three research questions:

1. How do people engage with different aspects of remote

USENIX Symposium on Usable Privacy and Security (SOUPS) 2021.
August 8–10, 2021, Virtual Conference.

communications in each context during the pandemic?

2. How do privacy and security factor into people's behaviors and attitudes towards aspects of remote communications in each context?
3. What approaches can be used to effectively inform and empower users' privacy and security decision making related to remote communications in each context?

We also designed our survey to allow us to explore related research questions for two other technologies that we hypothesized people would interact with more and/or have a new relationship with during the pandemic stay-at-home orders, namely smart home devices and social media platforms. Upon analyzing our results, we found that most participants' privacy and security concerns toward these two technologies did not change during the pandemic. Stay-at-home orders, however, significantly impacted participants' concerns and behaviors toward remote communications, which we primarily focus on in this paper. We include the survey questions on smart home devices and social media platforms and a summary of their findings in the extended version of our paper [18].

When being asked about conferencing tools, participants expressed a lack of decision-making agency. In WFH and LFH, participants reported to use the tool that was being decided for them by their employer or educator. Moreover, in all contexts, participants felt that they had no control over activating their webcam/microphone during their remote communications. For several participants, such imposed requirements contradicted their privacy and security preferences.

We found that participants' privacy attitudes and concerns towards the physical locations where their remote communications take place are context-dependent. By qualitatively analyzing participants' open-ended responses, we identified two types of location-related privacy: remote privacy (privacy from meeting attendees) and co-inhabitant privacy (privacy from household members). The open-ended responses suggested that in SFH, participants are mainly concerned about their co-inhabitant privacy, while valuing both remote and co-inhabitant privacy in WFH and LFH.

Based on the outcomes of our study, we distill several recommendations for organizations and tool developers on how to more effectively enable users to make informed and privacy-protective decisions with regard to their remote communications. In particular, we propose to enhance users' decision-making process by means of inclusive, transparent, and flexible policies on remote communications and designing privacy-protective features, which consider diverse and context-specific privacy and security needs.

2 Background and Related Work

Since December 2019, people around the world have been struggling with SARS-CoV-2 (novel coronavirus) and the resulting COVID-19 pandemic [19]. To help prevent further spread of the virus, many people have exercised social or physical distancing, i.e., keeping a safe distance from others who are not from the same household [20]. Consequently, people started working, socializing, and learning from home. As a result, the use of conferencing tools and audio and video communications has increased dramatically.

2.1 Privacy and Security Risks of Conferencing Tools

The pandemic has redefined home from a place of *privacy and security* [21] to a shared work, socializing, and learning space. This sudden shift from in-person to remote interactions has led to an unprecedented increase in the use of remote communication tools [22]. Teleconferencing and video conferencing tools, such as Zoom [23], Microsoft Teams [24], Google Hangouts [25], and WebEx [26], have all seen a massive rise in usage thanks to people working, socializing, and learning from home.

As people started to increasingly rely on such tools for their daily communications, experts have become more concerned about the wide range of privacy and security risks these tools expose their users to [9–12]. A few of the reported concerns include Zoombombing [27], undisclosed data mining [28], and selling information to third parties [29]. By considering the context around remote communications, literature has discussed the privacy and security concerns involving remote health-related sessions [30, 31], educational communications [15, 32–34], attending online courses [35, 36], and work-related meetings [14, 37, 38].

Experts have provided several guidelines aiming to prevent the risks and mitigate the potential harms of conferencing tools [39–41]. Despite being valuable sources of information, these guidelines put the burden of protecting privacy and security mainly on the user. This is an unrealistic expectation due to several reasons. Confirming the literature [42], our findings showed that privacy and security aspects, although being important, are not always the number one priority when using and interacting with conferencing tools. Moreover, our qualitative findings suggested that due to their roles in their organizations, users often have limited power in making privacy-protective decisions, especially in work- and education-related contexts. In addition, the best practices reported in the current guidelines constitute a broad recipe, hoping that they apply to all users in all contexts of remote communications. From the literature, we already know that privacy is context-dependent [43].

2.2 Home Audio and Video Broadcasting

During the pandemic, people started to rely more and more on the microphones and webcams of their devices to stay in touch with their colleagues, friends and family members, or their classmates. Only a few weeks into the pandemic, the market saw a 179% jump in the sales of webcams [44], followed by

a supply shortage [45–47].

Privacy and security experts have indicated that webcams and microphones are susceptible to risks and vulnerabilities. Several reports showed how easily hackers take control of users' devices and activate their built-in webcam and microphone by exploiting the device vulnerabilities [48–53]. During the pandemic, in all contexts of remote communications, users are at an even higher risk of such hacking incidents as they are spending an increased amount of time using their webcams and microphones in different locations of their homes to remotely communicate with others [54]. Users might not be aware that their webcams and microphones are turned on as the LED indicator lights are not always effective [55] or they might have been deactivated by the attacker [56, 57].

To prevent hacking attacks from happening, experts frequently recommend users to cover their webcams and microphones when they are not being used [58, 59]. During the pandemic, however, users might not be able to diligently exercise this protective approach as many are encouraged or even forced to have their webcams and microphones on all the time. Employers are setting always-on webcam policies to encourage spontaneous chats among employees [60] and using surveillance tools to closely monitor the activities of their workforce [61, 62], in some cases even without users' knowledge [63]. Saying no to such surveillance is not always easy, especially during the pandemic with the heightened risk of unemployment due to potential retaliations [64].

Remote learning is not immune to such commonplace imposed surveillance as well. School-issued devices are not transparent about whether they spy on students by activating their webcams and microphones [65]. Some schools use proctoring software that enables access to the students' webcams and microphones during the exams [66, 67]. In addition, policies are in place forcing students to have daily audio and video interactions with their peers or teachers [68].

The aforementioned privacy-invasive webcam and microphone policies and surveillance practices allude to the ineffectiveness of the blanket and commonly referenced solutions with respect to the rising risks of these technologies. Designing privacy-protective tools and providing usable privacy and security guidelines for users require a deep understanding of users' decisions and behaviors. Our study contributes to the body of literature by providing novel empirical evidence, which highlights the significant impact of the context of remote communications, as well as the living conditions, on attitudes and privacy concerns related to remote communications.

3 Methods

We launched an online worldwide survey (n =220) on Prolific in May 2020. We initially recruited 230 participants and excluded 10 of them: 3 participants used the open-ended boxes to advertise a product and 7 participants provided other irrelevant responses to open-ended questions. We provide the complete list of survey questions in Appendix A, and we mention the question number in parenthesis (e.g., CQ1) when referring to each survey question in the remainder of this section. The study protocol was approved by our Institutional Review Board (IRB).

3.1 Participant Recruitment

We recruited prolific participants who were at least 18 years old. Because of the worldwide impact of the COVID-19 pandemic, we did not restrict our respondents to a specific region and instead, recruited participants from all around the world. The survey took on average 16 minutes to be completed, and we compensated each participant with US$5.

3.2 Survey Procedure

We started the survey by introducing our study to be about "technology use in the home during the Coronavirus (COVID-19) Pandemic." We then asked a few questions to obtain participants' consent to participate in our study (see Appendix A.1).

We asked questions on three contexts of remote communications: working from home (WFH), socializing from home (SFH), and learning from home (LFH). In the survey, we showed questions related to each context in a separate block. We randomized the order of these blocks to mitigate the potential order bias [69]. We asked similar questions in the three tested contexts and only changed how we referred to remote communications in each context. Specifically, in the contexts of WFH, SFH, and LFH, we referred to remote communications as "remote work-related meetings," "remote personal meetings with friends and family members," and "remote learning-related meetings," respectively.

3.2.1 Context-Specific Questions

To control for participants' familiarity with the contexts of remote communications, at the beginning of each context, we asked participants to specify whether they have experience with remote communications in that context (CQ1). We implemented a logic so that respondents could see the remaining questions of that context only if they reported to have experience with the context in question.

To better understand our participants' timeline for remote communications, we asked questions to capture when they started remote communications and how often they were engaged in remote communications before and during the pandemic (CQ2-4). In each context, we explored participants' attitudes, behaviors, and privacy concerns related to three aspects of remote communications: conferencing tools (CQ5-10), modes of remote communications (CQ11-14), and locations of remote communications (CQ15-19).

To understand what our participants were most concerned about in their remote communications, at the end of each context, we asked respondents to specify the incidents that happened to themselves or others that they perceived to be

concerning or awkward (CQ20-22).

3.2.2 Demographics and Home Settings

Finally, we asked questions to understand participants' demographic information, as well as their home settings (DH1-16). We placed the demographic questions at the end of the survey to minimize the possibility of stereotype threat [70–72].

3.3 Data Analysis

To analyze responses, we conducted qualitative and quantitative analyses.

3.3.1 Qualitative Analysis

The first author was the primary coder, who created the codebook for each open-ended question and kept it updated throughout the coding process. To analyze the data, we applied structural coding [73], which is a question-driven qualitative coding approach to categorize the interview data as well as open-ended survey responses [74]. The codebook consists of main and sub codes. The main codes are created from the topics of interest in the study. For example, we were interested in understanding what factors led participants to use specific conferencing tools during the pandemic. In the codebook, the main code we used to answer this research question was *reasons to use conferencing tools*, which was then divided into 11 sub-codes (e.g., *functionality*) and further divided into 8 sub-sub-codes (e.g., *convenience and accessibility*). After the codebook was created, the first two authors used the codebook to independently code all the open-ended responses. Authors had several meetings to go over the codebook and the coded responses and resolve the conflicts stemming from mismatched understandings of the codebook. After agreeing on the definitions used in the codebook, the first two authors re-coded all the responses. The final codebook consists of 11 main codes, 122 sub-codes, 54 sub-sub-codes, and 4 sub-sub-sub-codes. For each codebook, the Cohen's Kappa inter-coder agreement was calculated after the second round of coding. The average rate of agreement for all the codebooks was above 0.91, with a minimum of 0.88 and a maximum of 1. Based on the literature, Cohen's Kappa inter-coder agreement of over 0.75 is considered as "excellent" [75]. We provide the final codebooks in Appendix B.

3.3.2 Quantitative Analysis

We fit $M = 4$ Cumulative Link Mixed Models (CLMMs) with logit as the link function to our collected data in order to explain the dependent variables (DVs) we asked our participants about. In each model, the DV is a categorical variable that can take multiple *ordinal* values, each of which we refer to as a *response category*. For all models, we treated participants' demographic and home setting factors as control variables. We considered Akaike Information Criterion (AIC) as the goodness of fit for the models [76]. We only report on demographic factors that helped the model fit significantly better than the model without them. It is important to note that we did not include the interaction terms in the final regression models as they did not improve the model fit. For the m^{th} model, $m \in \{1,\dots,M\}$, we denote the number of possible response categories by J_m, and we denote the corresponding number of *observations*, i.e., the number of participants that answered the question corresponding to that model, by N_m. For the n^{th} observation, $n \in \{1,\dots,N_m\}$, we let Y_m^n denote the observed response category. As per the CLMM definition, for the m^{th} model, $m \in \{1,\dots,M\}$, the probability that the n^{th} observation, $n \in \{1,\dots,N_m\}$, falls in the j^{th} response category or below, $j \in \{1,\dots,J_m-1\}$, is modeled as

$$\text{logit}(\Pr(Y_m^n \le j)) = \alpha_{j|j+1} - u_{\texttt{participant}_n} - \sum_{i=1}^{I_m} \beta^n_{\text{IV}_{m,i}},$$

where $\alpha_{j|j+1}$ denotes the threshold parameter or cut-point between response categories j and $j+1$, and $u_{\texttt{participant}_n} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_u^2)$ denotes the random effect for the participant in the n^{th} observation. Moreover, $\{\beta^n_{\text{IV}_{m,i}}\}_{i=1}^{I_m}$ represent model coefficients corresponding to the I_m different independent variables (IVs) in the m^{th} model, each particular to the level that was reported in the n^{th} observation.

4 Results

We start this section by providing information on participants' demographics and timelines of remote communications. We then present findings on participants' behaviors and decisions related to three aspects of remote communications: conferencing tools, modes of remote communications, and locations of remote communications.

4.1 Participants

We recruited 230 participants (reduced to 220 after excluding invalid responses) on Prolific. Our participants were mainly from UK (31%), Poland (15%), and US (14%). 43% of our respondents were female and 57% were male. Most participants did not have a background in Information Technology fields (65%) and were 18-29 years old (62%). We provide details on participants' demographics, home settings, and timelines of remote communications in Appendix C. Except for questions on the consent form (see Appendix A.1), none of the survey questions required participants to provide an answer. For each finding, we specify the number of participants that answered the corresponding question.

4.1.1 Frequency of Remote Communications

When asked about the frequency of remote communications before the pandemic, responses suggested that participants had more experience with remote communications in the socializing context than work and learning contexts. During the pandemic, in the contexts of WFH and SFH, most participants (WFH: 150/220, SFH: 208/220) reported that they have been mostly having remote meetings and communications. In the

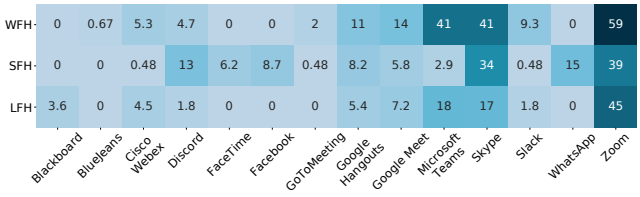


Figure 1: Usage of conferencing tools (in percentage) in different remote communication contexts reported by participants.

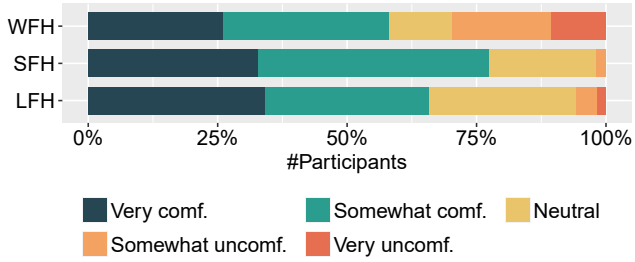


Figure 2: Participants' reported level of comfort with frequently used conferencing tools.

context of LFH, about half of our respondents (LFH: 114/220) reported to be having remote learning-related meetings.

Designing usable and privacy- and security-protective conferencing tools and guidelines in remote communications requires a deep understanding of users' attitudes and concerns towards remote communications. To this end, in our survey, we captured participants' context-specific thought process and decision making toward three aspects of remote communications during the pandemic: conferencing tools, modes of remote communications (webcam/microphone), and locations of remote communications. Without priming participants, we surfaced the role of privacy and security in participants' decision making related to each aspect of remote communications.

4.2 Conferencing Tools

In all contexts, Zoom was reported to be used more frequently than other tools (WFH: 35%, SFH: 27%, and LFH: 42%). We found that most participants (59%) were using the same conferencing tool for their WFH and LFH meetings and 35% of participants were using the same application in all three contexts. Figure 1 shows the fraction of participants who reported using each of the conferencing tools at least once for their remote communications across the three contexts.

In each context, on a five-point Likert scale, we asked participants to specify their level of comfort with the conferencing tool they most frequently use for their remote communications. Across all contexts, most participants (WFH: 87/150, SFH: 161/208, LFH: 75/114) were somewhat or very comfortable when using the conferencing tools for remote communications (see Figure 2). Our regression analysis indicated that the context of remote communications significantly impacts participants' level of comfort (see Table 1). We found that compared to work from home, participants were significantly more comfortable when using tools to communicate with family and friends (estimate $= 1.43$, p-value < 0.05) as well as communicating in the context of learning (estimate $= 0.88$, p-value < 0.05). Participants who reported to be using Google Hangouts were significantly more comfortable (estimate $= 2.13$, p-value < 0.05) with their conferencing tool than those who were using Zoom.

In each context, we explored participants' reasons behind their choice of conferencing tools as well as their comfort and discomfort with the tools. By qualitatively coding their open-ended responses, we surfaced several factors impacting participants' decision making and comfort level toward the use of conferencing tools.

4.2.1 Lack of Autonomy in Decision Making

In the contexts of WFH and LFH, participants frequently (WFH: 70/146, LFH: 64/107) implied that they have no agency over choosing what conferencing tool to use for their meetings. This lack of control was due to the fact that the tool was being selected for participants by their employers or educators, sometimes despite their personal preferences.

In the WFH context, some participants (WFH: 21/70) reported that the required conferencing tool is aligned with privacy and security preferences and requirements of their employers. P16 reported to be using Microsoft Teams for their WFH meetings: "Work requires me to only use this tool. They say this is the most secure one out there." Similarly P177 discussed why their employer asked them to use Microsoft Teams for WFH meetings: "It is the only tool that the company has approved security wise on our network."

In LFH, such imposed decisions contradicted some participants' (LFH: 17/64) personal privacy and security preferences, especially when they were required to use Zoom for their learning-related meetings. P36, who reported to use Zoom for their remote learning-related meetings, said: "That is the tool our teacher has chosen for us. Although security is certainly a problem."

4.2.2 Usability and Features

In all three contexts, the provided features and the usability of the tool were the second most frequently mentioned reasons to use the tool (WFH: 41/146, SFH: 92/204, LFH: 21/107) and the most commonly reported factors to make participants comfortable when using the conferencing tool (WFH: 75/111, SFH: 84/154, LFH: 42/70). P9, who frequently uses Microsoft Teams for their work meetings, said: "I can clearly see every file that's been attached to our meetings, I can easily contact with others and the quality of voice and video is just perfect." Unlike SFH, one of the most desirable features in the contexts of WFH and LFH was the ability of the tool to function properly with large groups. P102, who was using Zoom for their work meetings, said: "It supports a bigger number of people

Model No. and AIC	Dependent Variable	Independent Variable	Levels	Estimate	Odds Ratio	Std. Err.	p-value
1 (AIC=271.58)	Tool comfort level	Context (baseline=WFH)	SFH	1.43	4.18	0.56	*
			LFH	0.88	2.41	0.44	*
		Tool (baseline=Zoom)	Google Hangouts	2.13	8.41	0.81	*
			Google Meet	1.79	5.99	0.76	0.30
			Microsoft Teams	0.98	2.66	0.76	0.20
			Skype	0.80	2.23	0.74	0.27
		#Adults	$\{1,2,\dots\}$	−1.03	0.36	0.52	*
2 (AIC=266.71)	Microphone usage	Context (baseline=WFH)	SFH	1.53	4.61	0.39	***
			LFH	−1.61	0.20	0.33	***
		#Children (7-13)	$\{0,1,2,\dots\}$	1.29	3.63	0.44	**
3 (AIC=322.47)	Webcam usage	Context (baseline=WFH)	SFH	1.66	5.26	0.34	***
			LFH	−1.49	0.22	0.34	***
		Age (baseline=18-29)	30-49	0.98	2.66	0.46	*
			50-64	2.71	15.03	2.05	0.19
		#Rooms	$\{0,1,2,\dots\}$	0.31	1.36	0.14	*
4 (AIC=349.44)	Location comfort level	Context (baseline=WFH)	SFH	1.62	5.05	0.42	***
			LFH	0.66	1.93	0.36	0.06
		Location (baseline=Bedroom)	Dining room	−0.59	0.55	0.46	0.44
			Kitchen	−0.63	0.53	0.39	0.52
			Living room	−1.10	0.33	0.45	*
			Work room	−0.38	0.68	0.54	0.56
		#Adults	$\{1,2,\dots\}$	−0.56	0.57	0.28	*
		Gender (baseline=Female)	Male	1.04	2.83	0.42	*

Note: $^{*}p<0.05$ $^{**}p<0.01$ $^{***}p<0.001$

Table 1: Regression results of the CLMMs we built to explain participants' attitudes and concerns toward various remote communication aspects. A positive estimate of a level of an independent variable implies inclination toward an increase in the dependent variable and vice versa.

to be in a call better than the other ones."

In the WFH and LFH contexts, almost all participants (WFH: 40/41, LFH: 19/21) who mentioned using a conferencing tool based on its convenience and usability, reported to personally benefit the most from these attributes in their remote communications. P194 reported to be using Microsoft Teams for their WFH meetings: "Microsoft Teams allows me to stay connected more easily."

Unlike WFH and LFH, in SFH, several participants reported to use a conferencing tool mainly due to its perceived ease of use and convenience for others on the call (e.g., family members or friends), especially those with limited familiarity with technology. P90, who was most frequently using Skype to communicate with family members, said: "Parents are not confident with tech, Skype was the easiest for them to set up."

Some participants (WFH: 11/146, SFH: 9/204, LFH: 14/107) discussed giving up their privacy and security due to the tools' provided features and convenience. Almost all participants who mentioned such trade-offs reported to be using Zoom for their remote communications. In the context of WFH, P176 said: "Zoom offers the best features and is easy to use. Although security is certainly a problem." Users' trade-off between privacy and security and provided convenience is a known behavior in the literature [42, 77, 78].

4.2.3 Familiarity with the Tool

The most mentioned reason in deciding what conferencing tool to use to communicate with friends and family members was how familiar the tool was to participants themselves and also others on the call (SFH: 94/204). P54 reported to use WhatsApp for their SFH meetings to accommodate their family members: "It's the one my family have already installed on their phones and know how to use." Familiarity with the tool was the third most frequently mentioned reason in the contexts of WFH and LFH (WFH: 29/146, LFH: 15/107). P84 reported to use Skype more frequently than other tools: "The people I am calling with use Skype more than anything. I would rather use Zoom instead." Familiarity was also the second most commonly mentioned contributor to participants' comfort with conferencing tools (WFH: 22/111, SFH: 48/154, LFH: 15/70). P12 reported why they are comfortable with using Skype for their SFH meetings: "I have been using Skype since I was a teenager, so I'm used to it and that makes me more comfortable when I'm talking to friends and family."

Unlike WFH and LFH, in SFH 33% of participants, who reported to value the familiarity with the tool the most, implied that such familiarity partially stemmed from using the tool in contexts other than socializing (e.g., work, learning). P10 reported to use Microsoft Teams for their SFH meetings: "My

school uses the same platform and it's easier to be on only one platform at the same time."

Participants' open-ended responses implied how familiarity with the tool impacted their privacy and security concerns. A few participants (WFH: 9/29, SFH: 17/94, LFH: 6/15) perceived a sense of safety when using the tool due to their prolonged experience with the tool. P30, who used Discord for their personal meetings, said: "I know it is very safe and reliable because I've been using it for the past 3 years." This finding confirms the role of familiarity with technology in reducing risk perception [79]. Besides, a few participants associated their privacy concerns with their familiarity with the tool. P70 discussed why they only use Discord for their SFH meetings: "I already had account on it and also I am not comfortable sharing my info with more companies."

4.2.4 Privacy and Security Factors

In all three contexts, the perceived privacy and security of the tool were the third most commonly mentioned factors in making participants comfortable when using the tool for their remote meetings (WFH: 22/111, SFH: 20/154, LFH: 11/70). In WFH, Microsoft Teams and in SFH and LFH, Zoom were most frequently praised for their privacy and security practices.

When discussing their comfort with conferencing tools, some participants did not mention a specific privacy or security practice that made them comfortable when using the tool and instead said: "It is secure," "It feels like a safe app," or "I have no privacy concerns." We qualitatively coded the open-ended responses and identified several privacy and security best practices and perceptions that were frequently reported across all contexts:

- *Information being encrypted (7)*: WhatsApp:3, Zoom:4
- *Trusted brand (6)*: Microsoft Teams:2, Google Meet:1, Zoom:2, Cisco:1
- *No reported risk on media (6)*: Microsoft Teams:3, Cisco:1, WhatsApp:1, Discord:1
- *Protection from unauthorized access (5)*: Microsoft Teams:2, Google Meet:2, WhatsApp:1
- *Ability to set password for meetings (2)*: Zoom:2
- *No information being stored (1)*: Google Meet:1

Although most participants were comfortable with using the tools for their remote communications, some participants reported being somewhat or very uncomfortable (WFH: 17/150, SFH: 16/208, LFH: 7/114). Privacy and security were frequently mentioned as the reasons for participants' discomfort when using the tools (WFH: 8/17, SFH: 5/16, LFH: 3/7). Below is the list of privacy and security practices and beliefs that made participants uncomfortable when using the conferencing tool:

- *Risks and vulnerabilities reported by the media (7):* Zoom:7
- *Personal space being exposed in the meeting (4):* Zoom:3, Google Meet:1
- *Data being sold to third parties (2):* Zoom:2
- *Amount of information being collected (2):* Google Duo:1, Skype:1

We asked participants to specify how they manage their reported discomfort with the conferencing tools. In WFH and LFH, participants reported to address their concerns and discomfort by sharing less with other meeting attendees (WFH: 6/17, LFH: 2/7). Limiting the exposure was both in terms of restricting the content that is being shared, as well as modifying the configuration of their tool or the camera on their computer to limit the exposure. P4 limited the content they shared in the meeting and said: "I try not to say anything that could be used badly."

Unlike WFH and LFH, The most commonly mentioned mitigation approach in SFH was limiting or avoiding the use of the tool (SFH: 9/16). P188, who reported using Google Hangouts, said: "I will uninstall it as soon as it is no longer needed."

Some participants reported to take no action when being uncomfortable when using the tools (WFH: 4/17, SFH: 5/16, LFH: 1/7), mainly due to not being in charge of selecting the tools, not knowing what privacy and security controls the tool offers, or believing they have nothing to hide. P50 discussed why they do not take any action to address their privacy concerns with Google Duo: "Lots of other people use it too, nothing likely concerning will happen about what information the app ... collected on me."

4.3 Modes of Remote Communications

In all contexts, participants reported to activate their microphones significantly more frequently (p-value < 0.001) than their webcams when having remote communications (see Figures 3 and 4). Our CLMM results showed that compared to WFH, participants turned on their webcams (estimate $= 1.66$, p-value < 0.001) and microphones (estimate $= 1.53$, p-value < 0.001) significantly more often when having remote personal meetings with friends and family members, and significantly less often (webcam usage: estimate $= -1.49$, p-value < 0.001; microphone usage: estimate $= -1.61$, p-value < 0.001) when having remote learning-related meetings (see Table 1).

4.3.1 Microphone/Webcam Misuse

Accidental exposures of audio and video can lead to privacy violations. When asking participants about awkward incidents that had happened to them or others, across all contexts, the misuse of microphone and webcam was mentioned in almost all reported incidents (WFH: 29/34, SFH: 17/18, LFH: 8/10). In these incidents, the microphone and/or webcam were capturing unintended footage of a meeting attendee without their awareness and in some cases, without the awareness of the

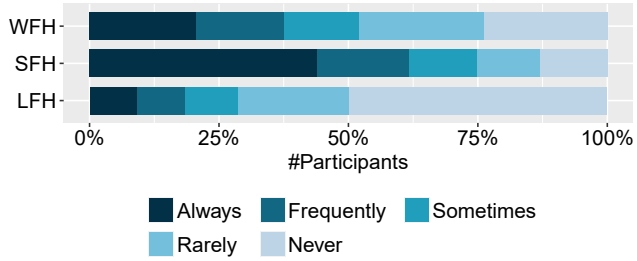


Figure 3: Reported frequency of using webcam.

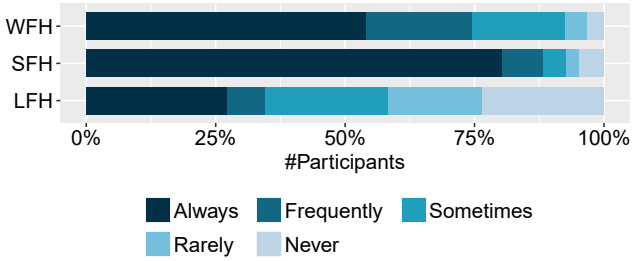


Figure 4: Reported frequency of using microphone.

household members. For example, in WFH, P194 mentioned an incident involving the microphone: "A member of management did not remember to mute himself while he answered his personal phone on speaker with what sounded to be a lawyer." P185 reported a similar incident in LFH that involved the misuse of webcam in the meeting: "Someone in a classroom stood up naked on the Zoom call and I guess he didn't know until it was too late."

In order to raise users' risk awareness and prevent such incidents from happening, we need to understand the underlying reasons for participants' preferences towards different modes of remote communications. We asked participants how they decide to turn on their webcam and microphone when having remote communications.

4.3.2 Agency over Decision Making

Participants' reasons to activate their microphone/webcam in WFH and LFH meetings implied their lack of agency over sharing their audio/video in their remote communications. In these contexts, respondents reported that they were explicitly expected to activate their microphone/webcam as a direct request by their employer or educator (WFH-Webcam: 73/101, LFH-Webcam: 66/94, WFH-Microphone: 77/127, LFH-Microphone: 59/97). Psychology literature refers to this type of behavior as *obedience*, i.e., a form of social influence where group members change their behaviors and attitudes due to a direct request or command from an authority figure [80]. P101, who reported to always turn on the webcam in their remote work-related meetings, said: "There isn't a choice in terms of my manager requesting a meeting face to face." For some participants, such imposed requests contradicted their personal preferences. P75 discussed their lack of desire to activate their webcam in work-related meetings: "If the manager asks me to turn on the video, I have to do it, but I personally prefer to maintain it switched off at all times."

In LFH, several participants reported being required to have their webcam and microphone on when taking exams (LFH-Webcam: 19/66, LFH-Microphone: 23/97). P43 discussed how they decide on when to activate their webcam during learning-related communications: "It depends if I'm explicitly asked by the professor to turn it on (For example when I have an 'oral exam' since written exams are now hard to do remotely)."

Although no participant reported to be explicitly requested by others to activate their microphone/webcam in SFH meetings, our qualitative analysis found the implicit expectation to be the main factor in participants' microphone/webcam usage. Most participants perceived lack of control over the use of microphone/webcam and reported that they are naturally expected to turn on their microphone/webcam when talking to their family and friends (SFH-Webcam: 134/208, SFH-Microphone: 166/208). P183 reported: "I always turn the webcam on during personal meetings with friends and family because I think it is what they expect and it would be rude not to."

Across all contexts, participants frequently mentioned that they would make the decision to activate their microphone/webcam based on other meeting attendees' behaviors (WFH-Webcam: 58/150, SFH-Webcam: 22/208, LFH-Webcam: 27/114, WFH-Microphone: 43/150, SFH-Microphone: 18/208, LFH-Microphone: 26/114). In LFH, P185 reported to sometimes turn on their webcam: "If other students have their cameras on I am more likely to turn mine on. But if nobody has theirs on, I will probably not turn on my camera." This type of social influence is called *informational conformity* [81], which serves as a cognitive repair [82] and happens when group members follow others' behaviors and directions as they are unsure about the appropriate behavior [83]. In several cases, participants reported to comply with the crowd despite holding a different preference. P43 reported to rarely turn on their webcam in work-related meetings: "Depending on the other person/people, and they always prefer to have a video. I personally find it a bit stressful, but don't mind it too much."

In SFH, some participants (SFH: 27/208) reported to jointly decide on the expectations around the use of microphone and webcam mostly prior to their personal meetings. P134 reported: "When we decide to meet, we choose video or none in the meeting invite." A few participants discussed the importance of joint decision making in accommodating meeting attendees' preferences. P160 reported to frequently turn on their webcam when meeting their friends: "If I miss seeing her face, we will plan a video call. We plan them because she has anxiety which I definitely want to accommodate for as

best [as I] can."

4.3.3 Attitudes over the Modes of Communication

Our participants shared diverse sentiments over activating their webcam/microphone in remote communications. In SFH, no participant suggested being uncomfortable with their lack of autonomy over webcam/microphone in their personal meetings. P117 discussed why they feel comfortable to always have their webcam on when meeting family and friends: "My family expect me to have video on, but I don't mind as I feel comfortable with people who really know me and accept me for who I am. I guess it's a gut feeling, if I don't feel anxious in their company in real life face to face, I would feel comfortable on a screen."

Unlike SFH, several participants (WFH: 21/101, LFH: 13/94) in WFH and LFH expressed negative attitudes toward having their webcam on. P80, who reported to activate their webcam at their employer's request, said: "I don't think video is necessary for the outcome of the meeting. It would be odd seeing colleagues in their home environment." P84 discussed why they do not feel comfortable with having webcam on in their LFH meetings: "I do not want to be seen by people I have never met, so I do not turn it on, unless I am being asked by the teacher." On the contrary, some participants (WFH: 16/101, LFH: 8/94) shared positive sentiments and supported having the webcam on during WFH and LFH meetings. P151, who reported to always turn on their webcam in WFH meetings, said: "I always turn it on as I feel face to face conversations with people create a better environment, and a higher level of honesty. I have campaigned for a policy in work to make video compulsory, and it has been taken up." P43 discussed why they preferred to have their webcam on in LFH meetings: "It is 'nice' and more productive during the Q&A meetings to have webcam on and discuss about issues/doubts about a particular project."

4.4 Locations of Remote Communications

We asked participants to specify which part(s) of their homes they most frequently use for their remote meetings. Across the three contexts, participants' bedroom (WFH: 56/150, SFH: 83/208, LFH: 49/114), living room (WFH: 32/150, SFH: 75/208, LFH: 25/114), and study or workroom (WFH: 38/150, SFH: 19/208, LFH: 21/114) were reported to be used more often than other locations. Despite being rare, a few participants reported using their bathrooms for their remote meetings (WFH: 3/150, SFH: 10/208, LFH: 1/114). Figure 5 shows the fraction of participants that reported to use each of the locations in their home at least once for their remote communications across the three contexts.

In all contexts, most participants (WFH: 123/150, SFH: 178/208, LFH: 91/114) were comfortable with the locations of their meetings (see Figure 6) while having significant differences across the contexts. The regression analysis showed that compared to WFH, participants were significantly more

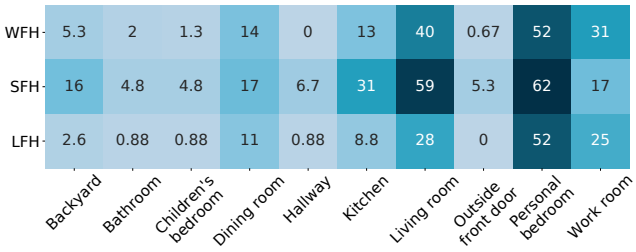


Figure 5: Usage of home locations (in percentage) reported by participants who engaged in remote communication contexts.

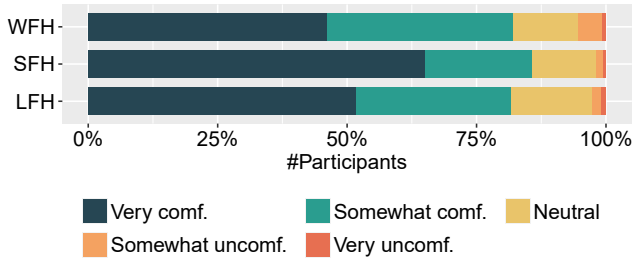


Figure 6: Participants' reported comfort level with frequently-used meeting locations.

comfortable when using any given location for their remote personal meetings (estimate $= 1.62$, p-value < 0.001).

4.4.1 Remote Privacy vs. Co-Inhabitant Privacy

In all three contexts, several participants reported to select a meeting location which they perceived to be the most private in their homes. Especially, in WFH, having privacy was the most frequently mentioned reason as to why a location is used for work meetings (WFH: 42/147). In the contexts of LFH and SFH, privacy was the second and third most common reason for participants' location-related decision making, respectively (LFH: 24/114, SFH: 35/208). Moreover, we found privacy and sense of safety to be frequently reported (WFH: 42/123, SFH: 68/178, LFH: 24/91) as participants' reasons to be comfortable with their meeting locations.

By qualitatively analyzing participants' open-ended responses, we identified two types of privacy: remote privacy and co-inhabitant privacy. Remote privacy refers to having privacy from other meeting attendees, while co-inhabitant privacy refers to having privacy from other household members. The types of privacy that were mentioned by participants varied among different contexts of remote communications.

In SFH, participants reported to have no concern over remote privacy as they felt comfortable with other meeting attendees (e.g., friends, family members) viewing their personal space. For example, P32 reported why they feel comfortable holding their personal meetings in the living room: "It's my living room, it's organized and everyone I talk to already knows it."

Almost all participants who considered privacy when se-

lecting their SFH meeting locations reported to be concerned with their co-inhabitant privacy. These participants reported to choose their meeting locations to be a personal space in their home where they are not being disturbed or interrupted by others during their remote meetings. Participants' personal space was mostly reported to be their bedroom. A few participants reported to have co-inhabitant privacy in other locations, including living room, kitchen, and workroom. P185 reported to use their bedroom for SFH meetings: "I feel the most comfortable in my own bedroom and I know that it will be a private space and that the least amount of interruptions will happen in my bedroom compared to other areas of the house." P25 reported why they were using the living room for personal meetings: "It is separated from the room in which my housemate works and so less likely that he will overhear."

Unlike SFH, in the contexts of WFH and LFH, participants' perception of privacy was more diverse. Some participants (WFH: 19/42, LFH: 15/24) reported to prefer having co-inhabitant privacy by detaching their meeting locations from other household members. P47 discussed why they decided to use their bedroom for remote work-related meetings: "I prefer using mostly my bedroom for most of my online work/meetings as I don't like others hearing me talk, I like to have a little bit of 'privacy'." P43 reported using their bedroom to preserve their co-inhabitant privacy: "I like to have my own little space, a bit of privacy from the rest of family and less distractions around so I can focus on the course." On the other hand, some participants were uncomfortable about others on the call seeing their personal space. These participants reported to desire having remote privacy by selecting a less personal location that provides a "neutral" or "professional" background with fewer details about their personal life. P195 talked about having privacy when holding their work meetings in the study or workroom: "I don't have anything private there that I would be unprofessional if I had to share my webcam with others." P122 discussed why they were using the kitchen for remote learning meetings: "This is the least personal place in the house to have such meetings."

Open-ended responses revealed potential conflicts between remote privacy and co-inhabitant privacy. In all contexts, participants who valued their co-inhabitant privacy frequently reported to be using their personal bedroom to have privacy from their household members. At the same time, participants in WFH and LFH perceived their personal bedrooms to be intimate and, therefore, not appropriate to preserve their remote privacy. We found similar tensions with regard to using living room for remote communications. Some participants chose to hold their remote meetings in the living room to have remote privacy, although having less co-inhabitant privacy due to the interruptions by household members.

4.4.2 Room Convenience and Equipment

Similar to participants' attitudes toward conferencing tools, convenience and comfort were frequently mentioned as the deciding factor when selecting a meeting location (WFH: 26/147, SFH: 65/208, LFH: 24/114). Especially, in the context of SFH, participants reported that the convenience of the location is the most important reason when choosing the room for their remote personal meetings. P80 discussed why they use the living room for their personal meetings: "This is the location of relaxation and the area where my husband and I can sit comfortably and talk to friends and family."

Another commonly mentioned reason behind participants' choice of meeting location in all three contexts was the presence of equipment that was needed for remote communications, including computer, desk, and books (WFH: 42/147, SFH: 41/208, LFH: 37/114). For participants in the context of LFH, the room equipment was the main factor in deciding what room to use for their learning meetings. P75, who reported to use their bedroom for LFH meetings, said: "This is the location where I have my desk and my PC in."

4.4.3 Discomfort with Meeting Locations

Although most participants were comfortable with their selected meeting locations, some respondents reported to be somewhat or very uncomfortable (WFH: 9/150, SFH: 4/208, LFH: 3/114). Across all contexts, the main factor participants mentioned that made them uncomfortable with a location was the perceived invasion of remote/co-inhabitant privacy when holding meetings there (WFH: 4/9, SFH: 1/4, LFH: 2/3). In the LFH context, P115 reported that they are uncomfortable with using their bedroom for remote learning-related meetings: "It is hard to get comfortable in the bedroom as it feels like a private area to invite people in to." P4 discussed their discomfort with using the living room for their remote work meetings: "I could be overheard and am not comfortable with the webcam being on as it intrudes on my privacy."

The open-ended responses indicated that only participants in the WFH context took steps to mitigate their discomfort with the location of their meetings, while in the contexts of SFH and LFH, participants reported to take no action when being uncomfortable with their remote meeting locations. The primary approach participants mentioned to take in the context of WFH was to limit the information exposure, either to other meeting attendees or their household members (WFH: 3/9). P38, who reported to mainly use their living room for their work-related meetings, said: "I minimise what can be seen and test the audio quality before the meeting." Similarly, P35 limited the work-related information from the household members: "I close my door and ask other family members [not to] come to the living room when having work meetings."

5 Discussion

We first provide a brief comparison between the contexts of remote communications. Based on our findings, we then discuss methods to inform and enable users' privacy-protective decision making related to remote communications.

5.1 Context-Specific Privacy Concerns and Attributes

We focused on three remote communication contexts: working from home (WFH), socializing from home (SFH), and learning from home (LFH). In each context, we surfaced participants' attitudes and concerns toward the use of remote communication technologies. Our quantitative and qualitative findings suggested several similarities and differences in participants' attitudes, behaviors, and privacy concerns among the three contexts. In all contexts, comfort and discomfort with conferencing tools and meeting locations were mainly explained by participants' privacy and security concerns and their perceived sense of safety. Our findings indicated that WFH and LFH were similar in terms of the choice and the use of conferencing tools (e.g., activating webcam/microphone). In SFH, unlike other contexts, the decisions toward the conferencing tools and the meeting locations were primarily based on the provided convenience.

Numerous articles have been published that provide recommendations on how to better protect privacy when engaging in remote communications [39, 41, 84]. Almost all of these guidelines are targeted toward the users, who are already struggling with an insurmountable mental pressure thanks to the pandemic. When an awkward incident happens in a conference call, end users are not the only group to blame, as they are only a small part of the remote communication ecosystem. Tool developers and users' employers and educators could play a critical role in informing and empowering users to adopt privacy-protective behaviors while communicating with others online.

The pandemic may not last forever, but remote communications will stay longer [16] and that requires us to critically examine what we have learned during the pandemic. Based on our findings, in the following, we distill several recommendations to inform and empower users, and to design more privacy-protective tools.

5.2 Enabling Context-Specific Informed Decision Making

Participants' open-ended responses showed lack of autonomy in their attitudes and behaviors toward remote communication technologies. Several participants reported to have no control over the choice of conferencing tools for their WFH and LFH meetings (see Section 4.2.1). Lack of active decision making was also apparent in participants' attitudes toward the use of webcam and microphone. Participants reported to be explicitly (WFH and LFH) or implicitly (SFH) expected to turn on/off their webcam/camera in the meetings (see Section 4.3.2). The qualitative findings indicated that having limited or no control over the conferencing tools and their features (e.g., webcam/microphone) was participants' primary impediment to managing their tool-related privacy and security concerns (see Section 4.2.4). To enable active and informed decision making in remote communications, we need to consider the context of the meeting.

Our findings suggested that WFH and LFH meetings have similar power dynamics that are being set by an authority figure (e.g., employer, educator). By providing **inclusive, transparent, and flexible policies**, workplaces and education institutes can take the first step toward informing and empowering meeting participants. To be inclusive, policies should acknowledge users' diverse and context-specific privacy needs and attitudes. To provide holistic privacy-protective policies, future studies should be conducted to explore other stakeholders' perspectives of remote communications, including but not limited to, employers and teachers.

In light of our findings, organizational policies need to discuss the choice of conferencing tool, the use of microphone and webcam in the meetings, and the available user controls. In addition, the policies should be flexible and open for feedback to help meeting attendees discuss and manage their concerns and discomfort. Items to be outlined in such policies include:

- What conferencing tools should be used for the meetings and why?
- What privacy and security controls are provided by the tools?
- In what condition are users (not) required to use their microphone/webcam?
- How can users control their microphone/webcam in the communication tools?
- How can meeting participants manage their concerns and discomfort with the tools?

Compared to WFH and LFH, in the context of SFH, participants felt being more in control of choosing a conferencing tool, which might be partially due to more balanced power dynamics. However, because of the implicit expectations, several participants felt having no control over the decision to activate their webcam/microphone in the personal meetings. As recommended by a few of our participants, **joint decision making** prior to the meeting could give meeting participants the opportunity to discuss their concerns and decide on a policy that accommodates and respects all of them.

5.3 Inclusive Privacy by Design

Across all contexts, the main factor participants mentioned to make them uncomfortable with a meeting location was the lack of remote and co-inhabitant privacy they felt when holding remote meetings in that location (see Section 4.4.1). Participants who referred to remote privacy reported that they do not feel comfortable having their home locations in the background of their WFH and LFH meetings. On the other hand, having a neutral or generic background was one of the frequently mentioned factors to make participants comfortable when using a meeting location (see Section 4.4.1).

Due to the restrictions posed by the diverse working, living, and learning arrangements, it may not be reasonable to ask everyone to find a neutral background for their remote meetings. Tool developers can enable features to help users protect their privacy. Some of the current communication tools, such as Zoom [85] and Microsoft Teams [86], already allow users to cover their real background by using virtual ones. Similarly, tools such as Skype [87] and Google Meet [88] provide a feature for users to blur their backgrounds.

Across all contexts, when discussing co-inhabitant privacy, several participants reported to be uncomfortable with other household members hearing their conversations (see Section 4.4.1). From the regression analysis, we found that an increase in the number of household members leads to a significant decrease in the level of comfort with conferencing tools as well as the locations of remote communications (see Table 1). To protect people's privacy in different contexts of remote communications, we need to design for diverse household settings. For example, to preserve co-inhabitant privacy in crowded settings, future remote communication devices can be enabled with a feature to detect and notify the user whether other household members are in the hearing range of their remote meetings. Such features can also respect the privacy needs of other meeting attendees, e.g., in case meeting participants are not comfortable with their voice or video being heard or seen by individuals who are not part of the call (e.g., household members).

5.4 Limitations

As the first paper to study remote communications at the transition of the pandemic, we surfaced participants' attitudes, behaviors, and concerns toward specific aspects of remote communications in different contexts. Due to the focus of our research and the survey methodology, we did not explore other potentially informative research questions, which could be studied in the future. In what follows, we will highlight the limitations of the current work, alongside several future research directions.

Our study used Prolific to recruit survey participants. Prior work recommended using Prolific to recruit a diverse sample of participants [89]. However, despite its diverse population and similarly to other crowdsourcing platforms, Prolific participants are not representative of any average population. For example, in Prolific, participants tend to be younger and more educated [90]. In addition, our participants were mainly from the UK, Poland, and the US, and we had a small number of participants from other countries (see Table 5). Due to these limitations, the findings of our study should not be generalized. Our study provides an overview of technology-related perceptions and behaviors during the global COVID-19 pandemic and we believe future studies can more directly focus on specific populations. In our study, participants' country of residence was not a statistically significant factor, which might be due to the small number of participants from some of the countries. Future studies could explore the difference in privacy concerns and attitudes among different countries and cultures.

As we previously mentioned, among other questions, our survey explored how participants' learning experience has been impacted by the COVID-19 pandemic. In this study, we only recruited participants who were at least 18 years old. However, it is also important to understand the impact of the pandemic and the privacy considerations of students from all ages, which should be considered in a future study.

To ensure participants' familiarity with the contexts of remote communications, for each context, we only asked the survey questions of participants who reported to have familiarity with that specific context. This potential selection bias might impact participants' attitudes and concerns toward remote communications in each context. Similarly, due to the nature of the job and depending on the level of experience, crowd-source participants might be more familiar with remote communication technologies than the average population. Having familiarity with a technology has been shown to decrease the amount of risk an individual would perceive related to that technology [79]. Therefore, the reported privacy and security concerns captured by our study could be lower than the average population' risk perception toward remote communications.

6 Conclusion

The COVID-19 pandemic has caused people around the world to abruptly shift their in-person work, personal life, and/or education meetings to remote ones, which could outlast the pandemic. Therefore, to enable safe remote experience, it is critical to design privacy-protective tools and empower users to consider privacy and security when engaging in remote communications. To this end, we conducted a 220-participant survey on Prolific, in which we considered three contexts of remote communications, namely working (WFH), socializing (SFH), and learning from home (LFH). Our quantitative and qualitative findings indicated that concerns, attitudes, and behaviors toward remote communications are diverse and context-dependent. Across all contexts, privacy and security were among the most frequently mentioned concerns that participants had. These concerns were exacerbated by the fact that participants felt that they had no agency over decision making about conferencing tools and the modes of remote communications. We provided several recommendations for tool developers and organizations to enable users to make privacy- and security-protective choices when engaging with remote communications.

Acknowledgments

We are especially grateful to our study participants. We thank Christine Geeng for reading the draft of our paper. We are also very thankful to our reviewers for their valuable feedback.

This work was supported in part by the NSF awards CNS-1565252 and CNS-1651230 and the University of Washington Tech Policy Lab, which receives support from the William and Flora Hewlett Foundation, the John D. and Catherine T. MacArthur Foundation, Microsoft, and the Pierre and Pamela Omidyar Fund at the Silicon Valley Community Foundation.

References

[1] Los Angeles Times, "Coronavirus social distancing around the world," https://www.latimes.com/world-nation/story/2020-04-06/coronavi-social-distancing-around-the-world, April 2020.

[2] R. Picheta, "Boris Johnson issues stay-at-home order, sending UK into lockdown to fight coronavirus pandemic," https://www.cnn.com/2020/03/23/uk/uk-coronavirus-lockdown-gbr-intl/index.html, accessed: 2020-10-28.

[3] M. Smith, "China forces millions of people to stay at home as virus toll rises," https://www.afr.com/world/asia/china-forces-millions-of-people-to-stay-at-home-as-virus-toll-rises-20200216-p5419e, accessed: 2020-10-28.

[4] The Washington Post, "Washington Post-ABC News poll March 22-25, 2020," https://www.washingtonpost.com/context/washington-post-abc-news-poll-march-22-25-2020/974c3312-5a40-4764-afb1-4bb6b86f1cf4/?itid=lk_inline_manual_2&itid=lk_inline_manual_2, accessed: 2020-10-28.

[5] D. Thompson, "The coronavirus is creating a huge, stressful experiment in working from home," https://www.theatlantic.com/ideas/archive/2020/03/coronavirus-creating-huge-stressful-experiment-working-home/607945/, March 2020.

[6] E. Koeze and N. Popper, "The virus changed the way we internet," https://www.nytimes.com/interactive/2020/04/07/technology/coronavirus-internet-use.html, April 2020.

[7] A. Tugend, P. Jordan, and M. Stein, "This school year has been unlike any other," https://www.nytimes.com/2020/10/14/education/learning/pandemic-school-remote-learning.html, October 2020.

[8] N. Daniels, "When the Pandemic Ends, Will School Change Forever?" https://www.nytimes.com/2020/05/05/learning/when-the-pandemic-ends-will-school-change-forever.html, May 2020.

[9] P. Wagenseil, "Zoom security issues: Here's everything that's gone wrong (so far)," https://www.tomsguide.com/news/zoom-security-privacy-woes, November 2020.

[10] Z. Doffman, "Why You Should Stop Sending Links On Facebook Messenger," https://www.forbes.com/sites/zakdoffman/2020/10/25/why-apple-iphone-and-google-android-users-should-stop-using-facebook-messenger-apps/?sh=1811fd9c33f1, accessed: 2020-10-28.

[11] C. Mihalcik, "Microsoft listened to Skype calls with 'no security' to protect recordings, report says," https://www.cnet.com/news/microsoft-listened-to-skype-calls-with-no-security-to-protect-recordings-report-says/, accessed: 2020-10-28.

[12] R. Perper, "WhatsApp disclosed 12 security flaws last year, including 7 classified as 'critical,' after Jeff Bezos phone was reportedly hacked," https://www.businessinsider.com/jeff-bezos-hack-whatsapp-disclosed-security-flaws-last-year-ft-2020-1, accessed: 2020-10-28.

[13] Zoom, "Attendee attention tracking," https://support.zoom.us/hc/en-us/articles/115000538083-Attendee-attention-tracking, accessed: 2020-10-28.

[14] B. Fung and A. Marquardt, "Millions of americans are suddenly working from home. that's a huge security risk," https://www.cnn.com/2020/03/20/tech/telework-security/index.html, March 2020.

[15] M. Liberman, "Massive shift to remote learning prompts big data privacy concerns," https://www.edweek.org/technology/massive-shift-to-remote-learning-prompts-big-data-privacy-concerns/2020/03, March 2020.

[16] M. Brenan, "U.S. Workers Discovering Affinity for Remote Work," https://news.gallup.com/poll/306695/workers-discovering-affinity-remote-work.aspx, April 2020.

[17] Prolific, "Quickly find research participants you can trust." https://www.prolific.co/, accessed: 2020-11-26.

[18] P. Emami-Naeini, T. Francisco, T. Kohno, and F. Roesner, "Understanding privacy attitudes and concerns towards remote communications during the COVID-19 pandemic," *arXiv preprint arXiv:2106.05227*, 2021.

[19] S. O'Grady, K. Bellware, M. Iati, L. Beachum, H. Denham, R. Thebault, and D. Sands, "The coronavirus has killed at least 1 million people worldwide," https://www.washingtonpost.com/nation/2020/09/28/coronavirus-covid-live-updates-us/, September 2020.

[20] Centers for Disease Control and Prevention (CDC), "Social Distancing," https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/social-distancing.html, accessed: 2020-10-28.

[21] D. G. Hayward, *Home as an environmental and psychological concept*, 1975.

[22] C. Trueman, "Pandemic leads to surge in video conferencing app downloads," https://www.computerworld.com/article/3535800/pandemic-leads-to-surge-in-video-conferencing-app-downloads.html, April 2020.

[23] J. Novet, "Why Zoom has become the darling of remote workers during the COVID-19 crisis," https://www.cnbc.com/2020/03/21/why-zoom-has-become-darling-of-remote-workers-amid-covid-19-outbreak.htmll, accessed: 2020-10-28.

[24] M. Hachman, "Microsoft's solution for COVID-19 is a free Teams subscription for six months," https://www.pcworld.com/article/3530374/microsofts-solution-for-covid-19-is-a-free-teams-subscription-for-six-months.html, accessed: 2020-10-28.

[25] I. Bonifacic, "Google makes Hangouts Meet features free in the wake of coronavirus," https://www.engadget.com/2020-03-03-google-makes-hangouts-meet-features-free-in-the-wake-of-coronavirus.html, accessed: 2020-10-28.

[26] J. Novet, "Cisco says Webex video-calling service is seeing record usage too, even as competitor Zoom draws all the attention," https://www.cnbc.com/2020/03/17/cisco-webex-sees-record-usage-during-coronavirus-expansion-like-zoom.html, accessed: 2020-10-28.

[27] T. Lorenz, "'zoombombing': When video conferences go wrong," https://www.nytimes.com/2020/03/20/style/zoombombing-zoom-trolling.html, accessed: 2020-10-28.

[28] A. Krolik and N. Singer, "A feature on Zoom secretly displayed data from people's LinkedIn profiles," https://www.nytimes.com/2020/04/02/technology/zoom-linkedin-data.html, accessed: 2020-10-28.

[29] J. Rosenblatt, "Zoom sued for allegedly illegally disclosing personal data," https://www.bloomberg.com/news/articles/2020-03-31/zoom-sued-for-allegedly-illegally-disclosing-personal-data, accessed: 2020-10-28.

[30] T. C. Li, "Privacy in pandemic: Law, technology, and public health in the covid-19 crisis," 2020.

[31] S. Bassan, "Data privacy considerations for telehealth consumers amid covid-19," *Journal of Law and the Biosciences*, vol. 7, no. 1, p. lsaa075, 2020.

[32] C. Venzke, "For remote learning, privacy challenges go beyond zoombombing," https://cdt.org/insights/for-remote-learning-privacy-challenges-go-beyond-zoombombing/, July 2020.

[33] S. Morrison and R. Heilweil, "How teachers are sacrificing student privacy to stop cheating," https://www.vox.com/recode/22175021/school-cheating-student-privacy-remote-learning, December 2020.

[34] J. Duball, "Shift to online learning ignites student privacy concerns," https://iapp.org/news/a/shift-to-online-learning-ignites-student-privacy-concerns/, April 2020.

[35] C. Caron, "How to protect your family's privacy during remote learning," https://www.nytimes.com/2020/08/20/parenting/online-school-privacy.html, August 2020.

[36] V. Strauss, "As schooling rapidly moves online across the country, concerns rise about student data privacy," https://www.washingtonpost.com/education/2020/03/20/schooling-rapidly-moves-online-across-country-concerns-rise-about-student-data-privacy/, March 2020.

[37] T. Spiggle, "Can employers monitor employees who work from home due to the coronavirus?" https://www.forbes.com/sites/tomspiggle/2020/05/21/can-employers-monitor-employees-who-work-from-home-due-to-the-coronavirus/?sh=3a31c6202fb7, May 2020.

[38] S. Morrison, "Just because you're working from home doesn't mean your boss isn't watching you," https://www.vox.com/recode/2020/4/2/21195584/coronavirus-remote-work-from-home-employee-monitoring, April 2020.

[39] V. Safronova, "Digital hygiene in the Zoom era," https://www.nytimes.com/2020/10/22/style/zoom-safety-protocols.html, October 2020.

[40] J. Engel Bromwich, "Protecting your digital life in 9 easy steps," https://www.nytimes.com/2016/11/17/technology/personaltech/encryption-privacy.html, November 2016.

[41] H. Drew, "How to protect your Zoom calls," https://www.washingtonpost.com/technology/2020/04/03/zoom-video-set-up/, April 2020.

[42] P. Emami-Naeini, H. Dixon, Y. Agarwal, and L. F. Cranor, "Exploring how privacy and security factor into iot device purchase behavior," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

[43] P. E. Naeini, S. Bhagavatula, H. Habib, M. Degeling, L. Bauer, L. F. Cranor, and N. Sadeh, "Privacy expectations and preferences in an iot world," in *Thirteenth Symposium on Usable Privacy and Security ({SOUPS} 2017)*, 2017, pp. 399–412.

[44] D. Howley, "Americans buying 'historic' amount of computers during coronavirus lockdown," https://finance.yahoo.com/news/americans-buying-historic-amount-computers-webcams-121043910.html, April 2020.

[45] C. Welch, "Webcams have become impossible to find, and prices are skyrocketing," https://www.theverge.com/2020/4/9/21199521/webcam-shortage-price-raise-logitech-razer-amazon-best-buy-ebay, April 2020.

[46] C. Baraniuk, "No end to covid-19 webcam shortage," https://www.bbc.com/news/technology-53506401, July 2020.

[47] B. Hill, "The covid-19 pandemic has caused a webcam shortage, and street prices are soaring," https://hothardware.com/news/covid-19-webcam-shortage-prices-skyrocketing, April 2020.

[48] J. Stern, "What i learned from the hacker who spied on me," https://www.wsj.com/articles/what-i-learned-from-the-hacker-who-spied-on-me-11549559728, February 2019.

[49] C. Williams, "Risk: Is this your webcam? you're being watched," https://www.wizcase.com/blog/webcam-security-research/, September 2019.

[50] M. Kuma, "Mac malware can secretly spy on your webcam and mic – here's how to stay safe," https://thehackernews.com/2016/10/macbook-camera-hacked.html, October 2016.

[51] K. O'Flaherty, "Zoom users beware: Here's how a flaw allows attackers to take over your mac microphone and webcam," https://www.forbes.com/sites/kateoflahertyuk/2020/04/01/zoom-users-beware-heres-how-a-flaw-allows-attackers-to-take-over-your-mac-microphone-and-webcam/?sh=7b2fe9e22fbe, April 2020.

[52] K. Johnson, "Ohio man charged for remotely controlling devices, spying on people for 13 years," https://www.usatoday.com/story/news/politics/2018/01/10/ohio-man-charged-remotely-controlling-devices-spying-people-13-years/1022275001/, January 2018.

[53] G. Kumparak, "Nasty facetime bug could allow others to eavesdrop on your microphone or camera," https://techcrunch.com/2019/01/28/nasty-facetime-bug-could-allows-others-to-eavesdrop-on-your-microphone-or-camera/, January 2019.

[54] H. Coffey, "Is it safe to leave your webcam uncovered after using video chat apps?" https://www.independent.co.uk/life-style/webcam-zoom-privacy-safety-tech-security-hack-working-home-lockdown-a9469156.html, April 2020.

[55] R. S. Portnoff, L. N. Lee, S. Egelman, P. Mishra, D. Leung, and D. Wagner, "Somebody's watching me? assessing the effectiveness of webcam indicator lights," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 1649–1658.

[56] M. Brocker and S. Checkoway, "iseeyou: Disabling the macbook webcam indicator {LED}," in *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, 2014, pp. 337–352.

[57] G. Cluley, "Webcam spying without turning on the led? researchers prove it's possible," https://grahamcluley.com/webcam-spying-without-turning-led-researchers-prove-possible/, December 2013.

[58] T. Germain, "How to protect yourself from camera and microphone hacking," https://www.consumerreports.org/privacy/how-to-protect-yourself-from-camera-and-microphone-hacking/, July 2019.

[59] D. Cook, "Zoom, skype, microsoft teams: Why you should cover the camera on your phone or laptop," https://nationalpost.com/news/canada/hackers-can-access-your-phone-and-laptop-cameras-heres-why-you-should-cover-them-now, April 2020.

[60] D. Safreno, "Loneliness is distracting," https://pragli.com/blog/loneliness-is-distracting/, March 2020.

[61] M. Jagannathan, "Like 'punching a time clock through your webcam': How employers are keeping tabs on remote workers during the pandemic," https://www.marketwatch.com/story/like-punching-a-time-clock-through-your-webcam-how-employers-are-keeping-tabs-on-remote-workers-during-the-pandemic-11596484344, August 2020.

[62] A. Holmes, "Employees at home are being photographed every 5 minutes by an always-on video service to ensure they're actually working — and the service is seeing a rapid expansion since the coronavirus outbreak," https://www.businessinsider.com/work-from-home-sneek-webcam-picture-5-minutes-monitor-video-2020-3, March 2020.

[63] D. West, "How employers use technology to surveil employees," https://www.brookings.edu/blog/techtank/2021/01/05/how-employers-use-technology-to-surveil-employees/, January 2021.

[64] P. Cohen and T. Hsu, "'rolling shock' as job losses mount even with reopenings," https://www.nytimes.com/2020/05/14/business/economy/coronavirus-unemployment-claims.html, June 2020.

[65] G. Genhart, "Spying on students: School-issued devices and student privacy," https://www.eff.org/wp/school-issued-devices-and-student-privacy, April 2017.

[66] D. Harwell, "Cheating-detection companies made millions during the pandemic. now students are fighting back." https://www.washingtonpost.com/technology/2020/11/12/test-monitoring-student-revolt/, November 2020.

[67] R. Deibert, "We've become dependent on a technological ecosystem that is highly invasive and prone to serial abuse," https://www.theglobeandmail.com/opinion/article-the-pandemic-has-made-us-even-more-dependent-on-a-highly-invasive/, November 2020.

[68] S. Johnson, "On or off? california schools weigh webcam concerns during distance learning," https://edsource.org/2020/on-or-off-california-schools-weigh-webcam-concerns-during-distance-learning/638984, August 2020.

[69] W. D. Perreault, "Controlling order-effect bias," *The Public Opinion Quarterly*, vol. 39, no. 4, pp. 544–551, 1975.

[70] T. Fernandez, A. Godwin, J. Doyle, D. Verdin, H. Boone, A. Kirn, L. Benson, and G. Potvin, "More comprehensive and inclusive approaches to demographic data collection," 2016.

[71] C. M. Steele, "A threat in the air: How stereotypes shape intellectual identity and performance." *American psychologist*, vol. 52, no. 6, p. 613, 1997.

[72] S. J. Spencer, C. M. Steele, and D. M. Quinn, "Stereotype threat and women's math performance," *Journal of experimental social psychology*, vol. 35, no. 1, pp. 4–28, 1999.

[73] J. Saldaña, *The coding manual for qualitative researchers*. Sage, 2015.

[74] E. Namey, G. Guest, L. Thairu, and L. Johnson, "Data reduction techniques for large qualitative data sets," *Handbook for team-based qualitative research*, vol. 2, no. 1, pp. 137–161, 2008.

[75] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.

[76] K. P. Burnham and D. R. Anderson, "Multimodel inference: understanding AIC and BIC in model selection," *Sociological methods & research*, vol. 33, no. 2, pp. 261–304, 2004.

[77] E. Zeng and F. Roesner, "Understanding and improving security and privacy in multi-user smart homes: a design exploration and in-home user study," in *28th {USENIX} Security Symposium ({USENIX} Security 19)*, 2019, pp. 159–176.

[78] P. Oppmann, "In digital world, we trade privacy for convenience," https://www.cnn.com/2010/TECH/04/14/oppmann.off.the.grid/index.html, April 2010.

[79] E. U. Weber, N. Siebenmorgen, and M. Weber, "Communicating asset risk: How name recognition and the format of historic volatility information affect risk perception and investment decisions," *Risk Analysis: An International Journal*, vol. 25, no. 3, pp. 597–609, 2005, accessed: 2020-06-03.

[80] R. B. Cialdini and N. J. Goldstein, "Social influence: Compliance and conformity," *Annu. Rev. Psychol.*, vol. 55, pp. 591–621, 2004.

[81] M. Deutsch and H. B. Gerard, "A study of normative and informational social influences upon individual judgment." *The journal of abnormal and social psychology*, vol. 51, no. 3, p. 629, 1955.

[82] C. Heath, R. P. Larrick, and J. Klayman, "Cognitive repairs: How organizational practices can compensate for individual shortcomings," in *Review of Organizational Behavior*. Citeseer, 1998.

[83] M. Sherif, "A study of some social factors in perception." *Archives of Psychology (Columbia University)*, 1935.

[84] B. Chen, "The Dos and Don'ts of Online Video Meetings," https://www.nytimes.com/2020/03/25/technology/personaltech/online-video-meetings-etiquette-virus.html, March 2020.

[85] Zoom, "Virtual Background," https://support.zoom.us/hc/en-us/articles/210707503-Virtual-Background, 2020, accessed: 2020-11-21.

[86] J. Spataro, "Custom backgrounds in Microsoft Teams make video meetings more fun, comfortable, and personal," https://www.microsoft.com/en-us/microsoft-365/blog/2020/06/12/custom-backgrounds-microsoft-teams-video-meetings-fun-comfortable-personal/, June 2020.

[87] Team Skype, "Introducing background blur in Skype," https://www.skype.com/en/blogs/2019-02-background-blur/, June 2019.

[88] Google, "Blur your background in Google Meet," https://workspaceupdates.googleblog.com/2020/09/blur-your-background-in-google-meet.html, September 2020.

[89] E. Peer, L. Brandimarte, S. Samat, and A. Acquisti, "Beyond the turk: Alternative platforms for crowdsourcing behavioral research," *Journal of Experimental Social Psychology*, vol. 70, pp. 153–163, 2017.

[90] Prolific Team, "What are the advantages and limitations of an online sample?" https://researcher-help.prolific.co/hc/en-gb/articles/360009501473-What-are-the-advantages-and-limitations-of-an-online-sample-, September 2018.

A Survey Questions

A.1 Informed Consent

This is a survey about technology use in the home during the Coronavirus (COVID-19) pandemic by researchers at the University of Washington, in Seattle, Washington, USA. The University of Washington's Human Subjects Division reviewed our study, and determined that it was exempt from federal human subjects regulation. We do not expect that this survey will put you at any risk for harm.

In order to participate, you must be at least 18 years old and able to complete the survey in English. We expect this survey will take about 20 minutes to complete. If you have any questions about this survey, you may email us at hometechnology@cs.washington.edu.

- I am 18 years or older.
  ○ Yes ○ No
- I have read and understand the information above.
  ○ Yes ○ No
- I want to participate in this research and continue with the task.
  ○ Yes ○ No

A.2 Context-Specific Questions (CQ)

In the contexts of WFH, SFH, and LFH, we referred to remote communications as "remote work-related meetings," "remote personal meetings with friends and family members," and "remote learning-related meetings," respectively. Here we only provide the questions for the WFH context.

- **CQ1:** Have you been mostly having remote work-related meetings from home during the COVID-19 pandemic?
  ○ Yes ○ No

  The rest of the context-related questions will only be presented if the answer is "Yes."
- **CQ2:** Before the COVID-19 pandemic, how often have you had remote work-related meetings from home?
  ○ Never ○ Once or twice a year ○ Once every 4-6 months ○ Once every 2-3 months ○ Once every month ○ Once every 2-3 weeks ○ Once every week ○ Not every day, but more than once a week ○ Every day
- **CQ3:** During the COVID-19 pandemic, how many hours a week do you spend in remote work-related meetings from home?
  ○ Less than 1 ○ 1 to 5 hours ○ 6 to 10 hours ○ 11 to 15 hours ○ 16 to 20 hours ○ 21 to 25 hours ○ 26 to 30 hours ○ 31 to 35 hours ○ 36 to 40 hours ○ Over 40 hours
- **CQ4:** During the COVID-19 pandemic, how long have you been having remote work-related meetings from home?
  ○ Since last week ○ Since two weeks ago ○ Since three weeks ago ○ Since one month ago ○ Since more than one month ago
- **CQ5:** During the COVID-19 pandemic, what conferencing tools do you mostly use for your remote work-related meetings? If you use more than one tool, please select the one you use most frequently.
  ○ BlueJeans ○ Google Hangouts ○ Google Meet ○ GoToMeeting ○ Microsoft Teams ○ Skype ○ Slack ○ UberConference ○ Zoom ○ Other (please specify [Open-ended])
- **CQ6:** Please explain why you have been using the tool that you have specified more frequently than other tools. [Open-ended]
- **CQ7:** During the COVID-19 pandemic, in your current environment, how do you feel about using this tool for your remote work-related meetings?
  ○ Very uncomfortable ○ Somewhat uncomfortable ○ Neither uncomfortable nor comfortable ○ Somewhat comfortable ○ Very comfortable
- **CQ8:** *(If in CQ7, Very uncomfortable or Somewhat uncomfortable is selected)* What about this tool makes you uncomfortable when using it? [Open-ended]
- **CQ9:** *If in CQ7, Very uncomfortable or Somewhat uncomfortable is selected)* How do you manage your discomfort when using this tool? [Open-ended]

- **CQ10:** *If in CQ7, Very comfortable or Somewhat comfortable is selected)* What about this tool makes you comfortable when using it? [Open-ended]

- **CQ11:** During the COVID-19 pandemic, how often do you turn on your device's webcam when having remote work-related meetings?
  ○ Never ○ Rarely ○ Sometimes ○ Frequently ○ Always

- **CQ12:** How do you decide whether or not to turn on your device's webcam when having remote work-related meetings? [Open-ended]

- **CQ13:** During the COVID-19 pandemic, how often do you turn on your device's microphone when having remote work-related meetings?
  ○ Never ○ Rarely ○ Sometimes ○ Frequently ○ Always

- **CQ14:** How do you decide whether or not to turn on your device's microphone when having remote work-related meetings? [Open-ended]

- **CQ15:** During the COVID-19 pandemic, which area of your home do you usually hold your remote work-related meetings in? If you use more than one location, please select the one you use most frequently for remote work-related remote meetings.
  ○ Backyard ○ Bathroom ○ Bedroom (yours) ○ Bedroom (your children's) ○ Dining room ○ Hallway ○ Kitchen ○ Living room ○ Outside front door ○ Study or workroom ○ Other (please specify [Open-ended])

- **CQ16:** During the COVID-19 pandemic, how do you feel about using this location to have remote work-related meetings?
  ○ Very uncomfortable ○ Somewhat uncomfortable ○ Neither uncomfortable nor comfortable ○ Somewhat comfortable ○ Very comfortable

- **CQ17:** *(If in CQ16, Very uncomfortable or Somewhat uncomfortable is selected)* What about this location makes you uncomfortable when having remote work-related meetings there? [Open-ended]

- **CQ18:** *If in CQ16, Very uncomfortable or Somewhat uncomfortable is selected)* How do you manage your discomfort when using this location for having remote work-related meetings? [Open-ended]

- **CQ19:** *If in CQ16, Very comfortable or Somewhat comfortable is selected)* What about this location makes you comfortable when having remote work-related meetings there? [Open-ended]

- **CQ20:** During the COVID-19 pandemic, have you or people you know ever experienced an awkward incident while having remote work-related meetings?
  ○ Yes ○ No

- **CQ21:** *(If in CQ18, Yes is selected)* Please describe the incident. [Open-ended]

- **CQ22:** *(If in CQ18, Yes is selected)* Please describe what you or people you know have done in response to the incident. [Open-ended]

A.3 Demographics and Home Settings

- **DH1:** Including yourself, how many adults 18 years of age and above live in your current home?
  ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ More than 5

- **DH2:** How many children at or above the age of 13 and under the age of 18 live in your current home?
  ○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ More than 5

- **DH3:** How many children at or above the age of 7 and under the age of 13 live in your current home?
  ○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ More than 5

- **DH4:** How many children under the age of 7 live in your current home?
  ○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ More than 5

- **DH5:** Who do you share your home with? (check as many as apply)
  ○ No one ○ Roommate(s) ○ Spouse(s)/Domestic partner(s) ○ Children ○ Parent(s) ○ Other (please specify [Open-ended])

- **DH6:** Do you have shared wall(s) with your neighbors?
  ○ Yes ○ No

- **DH7:** How many bedrooms does your home have?
  ○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ More than 5

- **DH8:** How many rooms other than bedrooms does your home have?
  ○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ More than 5

- **DH9:** What is your age?
  ○ 18-29 years old ○ 30-49 years old ○ 50-64 years old ○ 65 years and older

- **DH10:** What is your gender? [Open-ended]

- **DH11:** What is the highest degree you have earned?
  ○ No schooling completed ○ Nursery school ○ Grades 1 through 11 ○ 12th grade—no diploma ○ Regular high school diploma ○ GED or alternative credential ○ Some college credit, but less than 1 year of college ○ 1 or more years of college credit, no degree ○ Associates degree (for example: AA, AS) ○ Bachelor's

degree (for example: BA. BS) ○ Master's degree (for example: MA, MS, MEng, MEd, MSW, MBA) ○ Professional degree beyond bachelor's degree (for example: MD, DDS, DVM, LLB, JD) ○ Doctorate degree (for example: Ph.D., EdD)

- **DH12:** In which country do you currently reside? [List of countries provided by Qualtrics]
- **DH13:** What is your current employment status?
  ○ Full-time employment ○ Part-time employment ○ Unemployed ○ Self-employed ○ Home-maker ○ Student ○ Retired
- **DH14:** *(If in DH13, Unemployed or Retired is not selected)* The organization you work for is in which of the following?
  ○ Public sector (e.g., government) ○ Private sector (e.g., most businesses and individuals) ○ Non-for-profit sector
- **DH15:** Do you have a background in technology?
  ○ Yes ○ No
- **DH16:** *(If in DH14, Yes is selected)* Please specify what your technical background is. [Open-ended]

B Codebooks

The codebooks are available at:
https://gist.github.com/SOUPS-COVID-Privacy/97b6f6caeb13d5091314e6458049617d.

C Participants' Information

Timeline	Meeting Frequency	Context		
		WFH	SFH	LFH
Before the pandemic	Never	53%	33%	60%
	Once/twice a year	6%	7%	14%
	Once every 4-6 months	4%	6%	5%
	Once every 2-3 months	4%	7%	3%
	Once every month	3%	6%	2%
	Once every 2-3 weeks	4%	10%	2%
	Once every week	6%	13%	3%
	> once a week	11%	12%	5%
	Every day	9%	6%	6%
During the pandemic	< 1 (hour/week)	25%	35%	11%
	1-5 (hour/week)	44%	41%	40%
	6-10 (hour/week)	13%	14%	16%
	11-15 (hour/week)	7%	3%	12%
	16-20 (hour/week)	6%	2%	11%
	21-25 (hour/week)	1%	1%	7%
	26-30 (hour/week)	1%	1%	3%
	31-35 (hour/week)	2%	1%	0%
	36-40 (hour/week)	1%	1%	0%
	> 40 (hour/week)	0%	1%	0%

Table 2: Frequency of engaging in remote communications.

Experience Duration	Context		
	WFH	SFH	LFH
Since last week	1%	1%	7%
Since two weeks ago	0%	1%	5%
Since three weeks ago	3%	2%	9%
Since one month ago	5%	3%	18%
Since more than one month ago	91%	93%	61%

Table 3: Summary statistics of how long participants were experiencing the three contexts under study.

Question	Responses						
Shared wall(s) with neighbors	Yes 54%	No 46%					
Housemates	No one 3%	Roommate(s) 3%	Spouse(s)/Domestic partner(s) 19%	Children 11%	Parent(s) 21%	Other: Siblings 25%	
#Adults 18+ years old	1 12%	2 45%	3 22%	4 17%	5 4%	More than 5 0%	
#Children between 13 and 18 years old	0 81%	1 12%	2 6%	3 1%	4 0%	5 0%	More than 5 0%
#Children between 7 and 13 years old	0 86%	1 14%	2 0%	3 0%	4 0%	5 0%	More than 5 X%
#Children under 7 years old	0 83%	1 12%	2 5%	3 0%	4 0%	5 0%	More than 5 0%
#Bedrooms	0 0%	1 14%	2 28%	3 41%	4 14%	5 3%	More than 5 0%
#Rooms other than bedrooms	0 4%	1 11%	2 17%	3 27%	4 26%	5 15%	More than 5 0%

Table 4: Breakdown of participants' home settings.

Age		Gender		Highest Degree		Country of Residence		Employment		Tech Background	
18-29	62%	Female	43%	No schooling completed	0%	UK	31%	Full-time	41%	Yes	35%
30-49	34%	Male	57%	Nursery school	0%	Poland	15%	Part-time	17%	No	65%
50-64	4%			Grades 1 through 11	2%	US	14%	Unemployed	6%		
				12^{th} grade—no diploma	3%	Italy	7%	Self-employed	8%		
				Regular high-school diploma	21%	Portugal	7%	Home-maker	3%		
				GED or alternative credential	0%	Spain	4%	Student	25%		
				Some college credit, $<$ 1 year of college	4%	Greece	3%	Retired	0%		
				1+ years of college credit, no degree	16%	Canada	2%	Public sector	23%		
				Associate's degree (e.g., AA, AS)	4%	Other	17%	Private sector	67%		
				Bachelor's degree (e.g., BA, BS)	37%			Non-profit sector	10%		
				Master's degree (e.g., MA, MS, MBA)	13%						
				Professional degree (e.g., MD, JD)	0%						
				Doctorate degree (e.g., Ph.D., EdD)	0%						

Table 5: Participants' demographic information. Only countries with at least 5 participants are listed.