USENIX Association

Proceedings of SOUPS 2017
Thirteenth Symposium on
Usable Privacy and Security

July 12–14, 2017
Santa Clara, CA, USA

Conference Organizers

General Chair
Mary Ellen Zurko

Technical Papers Co-Chairs
Sonia Chiasson, *Carleton University*
Matthew Smith, *University of Bonn*

Technical Papers Committee
Yasemin Acar, *CISPA, Saarland University*
Lujo Bauer, *Carnegie Mellon University*
Konstantin (Kosta) Beznosov, *University of British Columbia*
Joe Calandrino, *Federal Trade Commission*
Lynne Coventry, *Northumbria University*
Heather Crawford, *Florida Institute of Technology*
Sergej Dechand, *University of Bonn, Fraunhofer FKIE*
Serge Egelman, *University of California, Berkeley and International Computer Science Institute*
Sascha Fahl, *CISPA*
Alain Forget, *Google*
Julien Freudiger, *Apple*
Simson Garfinkel, *National Institute of Standards and Technology and Georgetown University*
Nathan Good, *Good Research*
Apu Kapadia, *Indiana University Bloomington*
Hassan Khan, *University of Waterloo*
Katharina Krombholz, *SBA Research*
Ilaria Liccardi, *Massachusetts Institute of Technology*
Janne Lindqvist, *Rutgers University*
Michelle Mazurek, *University of Maryland*
Heather Patterson, *Intel Labs and NYU Information Law Institute*
Emilee Rader, *Michigan State University*
Robert Reeder, *Google*
Jessica Staddon, *North Carolina State University*
Elizabeth Stobert, *ETH Zurich*
Mary Theofanos, *National Institute of Standards and Technology (NIST)*
Janice Tsai, *Microsoft*
Blase Ur, *University of Chicago*
Yang Wang, *Syracuse University*
Rick Wash, *Michigan State University*
Tara Whalen, *Google*
Allison Woodruff, *Google*
Mary Ellen Zurko

Invited Talks Chair
Robert Biddle, *Carleton University*

Lightning Talks and Demos Co-Chairs
Heather Crawford, *Florida Institute of Technology*
Elizabeth Stobert, *ETH Zurich*

Karat Award Chair
Dave Crocker, *Brandenburg InternetWorking*

Posters Co-Chairs
Michelle Mazurek, *University of Maryland, College Park*
Kent Seamons, *Brigham Young University*

Tutorials and Workshops Co-Chairs
Adam Aviv, *U.S. Naval Academy*
Florian Schaub, *University of Michigan*

Publicity Chair
Patrick Gage Kelley, *University of New Mexico*

Steering Committee
Lujo Bauer, *Carnegie Mellon University*
Konstantin Beznosov, *University of British Columbia*
Robert Biddle, *Carleton University*
Sonia Chiasson, *Carleton University*
Sunny Consolvo, *Google*
Lorrie Cranor, *Carnegie Mellon University*
Apu Kapadia, *Indiana University Bloomington*
Patrick Gage Kelley, *University of New Mexico*
Heather Richter Lipford, *University of North Carolina at Charlotte*
Stuart Schechter, *Microsoft Research*
Matthew Smith, *University of Bonn, Fraunhofer FKIE*
Mary Ellen Zurko

External Reviewers

Wei Bai
Natã Miccael Barbosa
David Barrera
Jessica Colnago
Anastasia Danilova
Anupam Das
Martin Degeling
Pardis Emani-Naeini
Julie Haney
Roberto Hoyle
Ismail Khoffi
Meghan McLean
Alena Naiakshina
Olgierd Pieczul
Elissa Redmiles
Josh Tan
Christian Tiefenau
Daniel Votipka
Huichuan Xia
Yaxing Yao

SOUPS 2017
Thirteenth Symposium on Usable Privacy and Security
Message from the Chairs

Welcome to SOUPS 2017!

We are a USENIX conference again this year (and we look forward to continuing as one in 2018). The conference is a product of the hard work of all the organizers, the SOUPS steering committee, and the USENIX staff. We thank each and every one of you for your contributions to SOUPS 2017.

Starting last year, with our transition to a conference body independent of CMU support, we instituted a new structure around the SOUPS Steering Committee and its officers. Anyone can be elected to the SOUPS Steering Committee, terms are three years. If you'd like to find out more about serving on the steering committee, contact any member. The Chair of the Steering Committee is the General Chair of SOUPS. Mez will be serving out her final year as both for SOUPS 2018. A Vice Chair will be elected at the steering committee meeting this year, to serve as Chair for SOUPS 2019 and 2020.

SOUPS 2017 includes Workshops, Tutorials, Technical Papers, Posters (with reception), Lightning Talks, Demos, a Keynote talk, yet another reception, and the ever-popular ice cream social.

We thank each of our sponsors for their support—NSF, Facebook, and Google. SOUPS would not be possible without sponsor support.

Please visit our web site to learn the results of the SOUPS 2017 awards—Distinguished Paper, IAPP SOUPS Privacy Award, Distinguished Poster, the John Karat Usable Privacy and Security Student Research Award, and the SOUPS Impact Paper Award.

After two years of co-locating with USENIX ATC, we are making the move to co-locate with USENIX Security. We're looking forward to the benefits this change will bring in terms of both community interactions and finances. See you next year, August 12–14, at the Baltimore Marriott Waterfront!

Mary Ellen Zurko, *Independent Consultant*
General Chair

Sonia Chiasson, *Carleton University*
Technical Papers Co-Chair

Matthew Smith, *University of Bonn, Fraunhofer FKIE*
Technical Papers Co-Chair

SOUPS 2017:
Thirteenth Symposium on Usable Privacy and Security

Contents

Behaviour

Attacks and Defense

Privacy

Diversify to Survive: Making Passwords Stronger with Adaptive Policies

Sean M. Segreti, William Melicher, Saranga Komanduri, Darya Melicher, Richard Shay, Blase Ur[†], Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Michelle L. Mazurek[‡] Carnegie Mellon University [†]University of Chicago [‡]University of Maryland
{ssegreti, billy, sarangak, rshay, lbauer, nicolasc, lorrie}@cmu.edu,
darya@cs.cmu.edu, [†]blase@uchicago.edu, [‡]mmazurek@umd.edu

ABSTRACT

Password-composition policies are intended to increase resistance to guessing attacks by requiring certain features (e.g., a minimum length and the inclusion of a digit). Sadly, they often result in users' passwords exhibiting new, yet still predictable, patterns. In this paper, we investigate the usability and security of *adaptive* password-composition policies, which dynamically change password requirements over time as users create new passwords. We conduct a 2,619-participant between-subjects online experiment to evaluate the strength and usability of passwords created with two adaptive password policies. We also design and test a feedback system that guides users to successfully create a password conforming to these policies. We find that a well-configured, structure-based adaptive password policy can significantly increase password strength with little to no decrease in usability. We discuss how system administrators can use these results to improve password diversity.

1. INTRODUCTION

Reports of compromised password databases have become increasingly common in recent years [4, 12, 26, 35, 43, 50]. Such breaches can have far-reaching implications as they allow attackers to perform offline hash cracking attacks with virtually unlimited time. Because people commonly reuse passwords across accounts [11, 14], a breach of one account can compromise other accounts [14, 21]. While computationally expensive password-hashing functions are available, they are not always practical to implement or may be implemented ineffectively, and do not completely remove the existence of easy to exploit patterns. For high-value accounts, it remains imperative that users choose passwords that are hard for attackers to guess.

Password-composition policies, such as requiring a minimum length and inclusion of special characters, are commonly used to discourage users from choosing weak passwords. The usability and security of password-composition policies has been studied in depth [32, 45]; however, even under strict requirements, passwords still often have predictable patterns [23, 24, 45].

To increase a password set's resistance to guessing attacks, rather than focusing on the strength of individual passwords, researchers have proposed *adaptive* password-composition policies, which automatically evolve over time to encourage password diversity [34, 42]. For example, once some number of users of a given system have created passwords fitting a specific pattern, that pattern is banned and subsequent users may not create passwords fitting that pattern [34, 42]. While these proposed adaptive password systems may have strong potential benefits for security, their impact on password strength and usability has yet to be empirically tested.

In this work, we evaluate the security and usability impact of making password-composition policies adaptive. We focus on two implementations of this approach that do not require storing a copy of the plaintext (or reversibly encrypted plaintext) passwords, and which can operate with a traditional (non-adaptive) password policy. Adaptive policy systems that store plaintext or reversibly encrypted passwords are insecure in real-world situations, where one must assume attackers may gain access to the password store.

Our primary focus, Leininger et al.'s PathWell [33, 34], prohibits users from creating passwords with the same character-class *structure* (pattern of symbols, digits, and letters) as another user's password. When a new password is created, its structure is deemed "in use" and is not allowed during future password creation attempts. To increase the usability of the PathWell structure-based approach, we designed and tested a feedback system that guides users to choose a password with a permitted structure. The second approach, introduced by Schechter et al. [42], instead uses a specialized Bloom filter to probabilistically prevent users from creating passwords that are deemed too popular.

To evaluate the security and usability of these approaches, we conducted a two-part, between-subjects online study. 2,619 participants created a password under one of twelve conditions, designed to study: how adding adaptive requirements to traditional password policies affect security and usability; whether participants are confused by the extra requirements; how security and usability change as the stringency of adaptive policies increase; and the effect of graphical feedback.

We found that the passwords created under structure-based adaptive password-composition policies can be several or-

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

ders of magnitude more secure than those created without an adaptive password-composition policy. More surprisingly, we found that, structure-based adaptive policies can be applied without a significant usability cost, according to numerous usability metrics. We observed no statistically significant differences in creation time, password recall, or password storage (how often passwords were written down) between pairs of conditions that differed only in whether an adaptive policy was used. The only noteworthy usability downside of applying structure-based policies was that participants needed (on average 0.58–1.58) more attempts to create their passwords; however, this neither significantly impacted the overall time to create a password (of which a single attempt is a small fraction) nor affected user sentiment, except for the condition which simulated the most extreme numbers of disallowed structures. Our attempts to provide additional feedback to overcome the expected usability penalty of structure-based adaptive policies were largely superfluous; little usability had been lost to begin with.

2. BACKGROUND AND RELATED WORK

We first discuss the types of password-guessing attacks that adaptive policies aim to mitigate. We then detail the manner in which password-guessing approaches exploit common patterns in the absence of adaptive policies. Finally, we discuss related work on password-composition policies.

2.1 Password-Guessing Attacks

Password-guessing attacks fall broadly into one of two categories: those for which guessing is limited to a relatively small number of attempts, and those for which large-scale guessing is possible. An example of the former category is an online attack, in which an attacker submits password guesses to a running system. Because a well-configured system will have a policy that rate-limits guessing or locks accounts following a small number of incorrect authentication attempts, attackers are limited to making some of the most likely guesses. Measurements of fraudulent SSH login attempts revealed that some of the most common passwords that attackers guess are passwords often found in data breaches, as well as passwords related to system administration (e.g., variants of "root") [1]. If the adaptive system is bootstrapped, these types of common passwords could be banned initially by an adaptive policy. If not, the threat of password guessing is minimized to only the small number of accounts permitted to pick such a password before that password is banned.

Large-scale guessing attacks also present a major threat in a number of different circumstances. One such situation is an offline attack aimed at discovering credentials reused from other sites. If an attacker obtains a store of hashed passwords, which has become unfortunately common in recent years [4, 12, 26, 35, 43, 50], the attacker can perform offline hash cracking, limited only by his or her time and resources. Because users often reuse passwords across accounts [11,14], attackers can use the credentials obtained in an offline attack to compromise other accounts [14, 21].

Password-specific hash functions are designed to be computationally expensive in order to limit the number of guesses an attacker can make. Unfortunately, their deployment has proven error-prone in practice [25], is difficult to implement on some popular platforms that support backwards compatibility with legacy systems [38], and does not remove the existence of some easily exploitable patterns. Researchers have proposed systems to prevent offline cracking attacks [29], though these systems have yet to be deployed in practice and rely on having accurate models of generating artificial, yet plausibly human-chosen, passwords.

Even if system administrators were to follow all best practices to prevent offline password cracking for web accounts, other situations in which passwords are used would still be vulnerable to offline guessing. Encrypted file containers and full-disk encryption, as well as password stores from password managers (encrypted with a key derived from a master password), would remain vulnerable to offline guessing if an attacker gains access to the relevant file or device. Because adaptive schemes would not adapt over time in these single-user systems, these schemes could be bootstrapped with likely password patterns.

Despite the ability for system administrators to rate-limit online attacks and employ some technical mechanisms to minimize, but not completely eliminate, the threat of offline attacks, a user concerned about his or her high-value accounts is incentivized to practice defense in depth. Rather than relying exclusively on a system administrator to follow all best practices perfectly, which is far from guaranteed in practice, a user should choose unique passwords that are hard for attackers to guess. As we show in this paper, adaptive policies better enable users to do so.

2.2 Guessing Common Patterns

The types of common password characteristics adaptive policies aim to avoid can be exploited by password-cracking approaches. For example, Weir et al. proposed a probabilistic context-free grammar (PCFG) to model passwords [53]. Based on training data of previously observed passwords, PCFG assigns probabilities to both password structure (e.g., *princess111* has the structure {*8 letters*}{*3 digits*}) and constituent strings (e.g., "111"). Kelley et al. proposed improvements to this method [30], e.g., to treat uppercase and lowercase letters independently. Other researchers have advocated using grammatical structures and semantic tokens as non-terminals [40, 51]. Komanduri recently offered several PCFG improvements, including string tokenization and assigning probabilities to terminal strings not seen in training data [31]. The PCFG and its variants have been used in a number of prior studies to gauge password strength [10, 13, 16, 30, 36, 37, 45, 48]. Structure-based adaptive policies [34] make the PCFG approach less effective because the PCFG relies on the commonality of password structures to guess likely passwords.

Markov models also effectively model human-chosen passwords. Narayanan and Shmatikov first proposed using a Markov model of letters in natural language with finite automata representing password structures [39]. Castelluccia et al. used a similar method as part of their password meters [8]. Recently, Dürmuth et al. [15] and Ma et al. [36] evaluated the effectiveness of multiple variations of Markov models for cracking passwords, finding that Markov models were more accurate than PCFG at guessing passwords under certain circumstances. Popular password-cracking software packages, such as John the Ripper[1] and Hashcat,[2] offer variants of a Markov model.

Adaptive policies have two conceptual advantages that aid in resisting guessing by Markov models. Structure-based adaptive policies encourage passwords with unpredictable structures, which are likely to foster character-level unpredictability that in turn may be hard to capture in a Markov model. Furthermore, string-based adaptive policies forbid the predictable passwords that a Markov model would easily guess.

Adaptive policies also provide conceptual difficulties for the guessing approaches of common password-cracking software tools. For example, the Hashcat toolkit implements a "mask attack," in which password guesses are generated by progressively exhausting the keyspace of each structure in an attacker-defined ordered list.[3] Despite its brute-force component, this approach can be effective in real-world cracking because many users craft passwords matching popular structures [41, 46].

2.3 Password-Composition Policies

Human-chosen secrets frequently share predictable, and thus exploitable, characteristics [5]. To discourage such patterns, organizations like the National Institute of Standards and Technology (NIST) recommend that system administrators employ password-composition policies, such as mandating a minimum length and the inclusion of a digit [7]. However, passwords created under these guidelines still frequently have exploitable patterns, such as consisting of a dictionary word followed by a number and symbol [37]. Furthermore, while some policies are better than others at balancing the tradeoff between leading users to create passwords that are harder to guess and improving the usability of password creation [45], particularly onerous password-composition requirements can unduly burden and annoy users [32].

Beyond requiring that passwords be at least a particular length and contain particular classes of characters, policies can also prohibit (blacklist) the most popular or predictable passwords. A judiciously chosen blacklist can lead users to pick passwords that are far harder to guess than those created without a blacklist [27, 32, 52]. String-based adaptive policies essentially build a blacklist that expands over time to reflect new password patterns.

Common patterns can make passwords easy to guess, yet they also can make them easy to remember. However, a recent study by Bonneau and Schechter showed that people are capable of remembering a large set of random characters if they are presented using spaced repetition [6]. Adaptive policies strive to capitalize on this discovery and, by introducing more complexity, to find a way to prompt users to create passwords with fewer exploitable patterns and thus higher resistance to guessing attacks.

3. METHODOLOGY

We conducted a two-part online study to examine how participants create and use passwords under two adaptive password policies in multiple configurations. In the first part of the study, we asked participants to create a password under a specific policy, take a survey, and then recall their password. Two days later, we asked participants to return and recall their password, in addition to completing a second survey. We recruited participants through Amazon's Mechanical Turk crowdsourcing service (MTurk). We required that participants be at least 18 years old and located in the United States. Our overall methodology is based on techniques used to compare password-composition policies in prior work [30, 32, 44, 45, 47, 48]. Our protocol was approved by our institution's IRB.

In *part one* of our study, we asked participants to imagine their main email account had been compromised, and they must create a new password. Prior work suggests that asking participants to imagine creating a password for their email account leads to stronger passwords than simply creating passwords for a study [30, 32]. We informed them that they would be asked to re-enter their password in a few days, and instructed them to do whatever they would normally do to remember and protect a new password.

We then showed participants one of twelve sets of password-creation instructions, depending on their assigned condition, described in Section 3.4. After creating a password, participants completed a survey on the password creation experience, as well as how they chose their password. We then asked participants to recall their password. Participants who typed their password incorrectly five times were then shown their password.

Two days later, we invited participants via email to return for *part two* of the study in which we asked participants to recall their password. After five incorrect attempts, participants were shown their password. Participants could follow a "Forgot Password" link to be emailed a link to their password. Next, we administered another survey about the steps the participant took to remember their password, including whether and how participants stored their passwords (e.g., writing it down or saving it electronically).

Our data-collection method enables us to measure several quantitative usability metrics during password creation and recall. We collect timing information and the number of password creation/recall failures. We use electronic copy-paste/autofill detection during the recall phase to augment the self-reported survey data. We also ask participants sentiment questions about the ease of both creating and recalling a password.

3.1 Adaptive Policies

We evaluated two adaptive password-policy systems. Both of these systems have two characteristics that we consider essential for an adaptive system. First, a secure system must not store passwords in plaintext or reversibly-encrypted ciphertext, as this creates a new avenue of attack. Second, an effective system must integrate with traditional password-composition policies for ease of deployment.

The first adaptive policy we evaluate operates on password structures, the password's sequence of character classes (uppercase, lowercase, digit, or symbol). We implement KoreLogic's Password Topology Histogram Wear-Leveling (PathWell) [33], which is designed to enforce password structure diversity, and refer to this system as the *structure-based ap-*

[1] 1 `www.openwall.com/john/`
[2] 2 `hashcat.net`
[3] 3 `hashcat.net/wiki/doku.php?id=mask_attack`

proach. In its simplest form, PathWell would require that all passwords in a set must have a unique structure. For example, if the password 'passWord11!' is in the set, then 'asdfQwer99#' would not be allowed for future passwords, because they both have the same character-class structure.

However, blacklisting a character-class structure after a single use could potentially help attackers by letting them know that, once they have successfully found a hash preimage, no other passwords in the set have the same character-class structure, obviating additional guesses within that structure. Additionally, it can decrease the usability of the system as more passwords are created. Therefore, PathWell also enables a structure to be blacklisted after some number of passwords use that structure, and this is the approach we use. PathWell's structure blacklist is designed to be preloaded with commonly used structures, and to grow over time as users are added to a system.

The second adaptive policy operates on passwords as a whole, rather than their character-class structure. In particular, we evaluate Microsoft Research's "Popularity is Everything" system, which prevents a given password from being used too many times in a system [42]. To do so, this approach uses a specialized database based on a Bloom filter [3] to record how many times a password is used without storing the password itself. As with PathWell, this database grows over time. We refer to this as the *string-based approach*.

3.2 Password-Composition Policies

To prevent users from creating passwords that are very weak or especially short (and vulnerable to brute-force attacks), adaptive policies should be used together with password composition requirements. Because early adopters of adaptive password policies in the real world are most likely organizations with high security needs, we chose to focus on stronger-than-average password-composition policies commonly employed in organizational and government settings, rather than for run-of-the-mill online accounts.

Historically, password-composition policies for higher security settings have mandated many different character classes. For example, in what we term the *4class8* (*4c8* for short) policy, passwords must contain at least eight characters, including all four character classes (lowercase letters, uppercase letters, digits, and symbols). This policy was recommended by NIST guidelines in 2011 [7] and once represented the *de facto* industry best practice. Such a policy is still popular, leading us to study it. Recently, however, password-composition policy guidance has begun to emphasize password length, rather than including character classes. Such a shift is evident both in the academic research literature [19, 45], as well as in the mass media [2]. U.S. government accounts have begun to deploy policies that emphasize password length [9].

We focus most of our experiments on a password-composition policy which we term *3class12* (*3c12* for short). This policy requires that passwords contain at least twelve characters, as well as three of the four character classes. This particular 3class12 policy has been found in prior work to better balance the security-usability tradeoff than policies like 4class8 [45] and is similar to policies in use on U.S. government systems [17].

		# Structures		# Passwords	
Size	PCFG	3c12	4c8	3c12	4c8
M	10^5	2,141	2,236	1.62E56	4.64E28
L	10^6	8,940	—	1.65E56	—
XL	10^7	48,199	—	4.39E59	—

Table 1: Description of the blacklists we used. The first two columns show the how many password guesses were modeled in generating that blacklist. The remaining columns describe how many structures and passwords each blacklist disallows for the 3c12 and the 4c8 polices.

3.3 Systematically Testing Adaptive Policies

Evaluating an adaptive policy experimentally poses a unique challenge: if participants are working with an adaptive policy on a real-world system, each participant's password will modify the blacklist, thereby creating a unique environment for each participant. Instead, we opted to use pre-calculated blacklists of different sizes, allowing us to collect results from hundreds of participants exposed to the exact same situation. As such, we effectively compare passwords created at different points in the adaptive process, and those without an adaptive policy.

For this evaluation to succeed, it is critical to build blacklists that capture the most popular passwords. We gathered a total of 32,965,921 passwords from public leaks [22, 43, 50]. However, relatively few passwords from these sets meet the requirements of our stringent baseline policies (Section 3.4), limiting the size of the blacklists we could generate. To compensate, we trained a PCFG guesser [30] with these leaks and used it to enumerate the most probable guesses that conform to the minimum requirements. Using these guesses, we computed blacklists of the most common structures and passwords. This process simulates initial users in an adaptive system choosing highly probable passwords, with the corresponding structures subsequently being blacklisted. In Table 1, we summarize the blacklists we evaluated. Note that the M blacklist serves two purposes: Its corresponding 2,141 unique character-class structures are used to configure the structure-based adaptive approach; the 10^5 passwords used to generate the blacklist are the passwords that are banned in the string-based approach.

3.4 Conditions and Research Questions

We assigned participants to one of twelve conditions, each with different requirements, instructions, and feedback. Because of the large number of possible factors, it was not feasible to test all combinations of factors in isolation. Instead, we chose to run a set of conditions spanning five research questions (RQs) detailed below.

RQ1: Impact of Structure-Based Adaptive Policy *How are the usability and security of passwords affected when using a structure-based adaptive policy in addition to a traditional policy?* To answer this question, we evaluate two traditional password-composition policies: 3c12 and 4c8. We test each policy both with and without a medium-sized blacklist of character-class structures. The following conditions address our first research question:

- **3c12**: Passwords must contain at least 12 characters and include at least three character classes. This policy has been recommended in the academic literature [45]

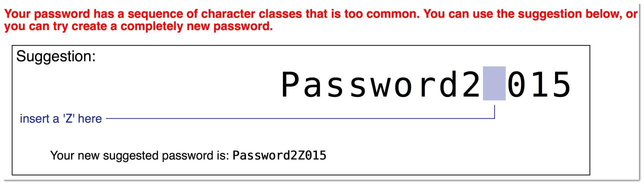


Figure 1: The interface shown to participants in standard structure-based blacklist conditions (Struct$_\mathrm{M}$, Struct$_\mathrm{L}$, Struct$_\mathrm{XL}$, Struct$_\mathrm{M}$Ins, Struct$_\mathrm{M}$Sub, and Struct$_\mathrm{M}$4c8) when their password was rejected by the structure blacklist. The suggested modification in this example was randomly chosen as insert a "Z" character. The Struct$_\mathrm{M}$3Hint interface is similar, but instead shows three different suggested modifications.

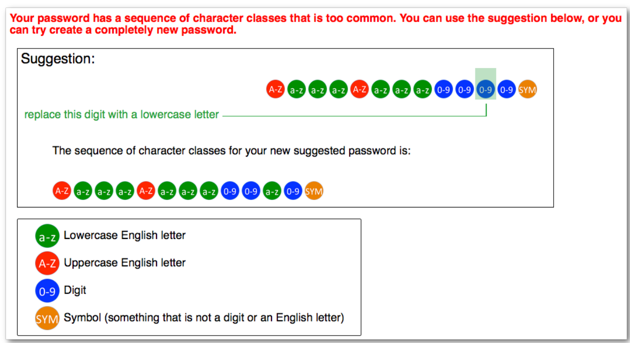


Figure 2: The interface shown to participants in Struct$_\mathrm{M}$S when their password was rejected by the structure blacklist. The interface shows the character-class structure of the attempted password, rather than the password itself. The Struct$_\mathrm{M}$SV interface was similar, except the structure was displayed in real time as it was typed. The Struct$_\mathrm{M}$Hyb interface was also similar, except only the inserted or substituted characters were obfuscated.

and is similar to policies for many U.S. government accounts [9] (e.g., [17]).

- **Struct$_\mathrm{M}$**: The 3c12 policy plus a structure-based adaptive policy with 2,141 banned structures (corresponding to the first 10^5 PCFG guesses, as shown in Table 1).
- **4c8**: Passwords must contain at least 8 characters and include all four character classes. This policy was recommended by NIST guidelines in 2011 [7] and once represented best practice. Note that this condition is unique in that data for it was collected as part of an earlier study that used the same data-collection methodology. The condition does not contribute to any of our main results, but we include it here as an additional, informative baseline.
- **Struct$_\mathrm{M}$4c8**: The 4c8 policy plus a structure-based adaptive policy with 2,236 banned structures, as shown in Table 1.

As part of this work, we designed and implemented a feedback system for the structure-based adaptive policies. If a participant's attempted password used a banned structure, the feedback system displayed a randomly selected modification (either insertion or substitution of a random character or character class) at a randomly selected location to their initial password such that the new password is guaranteed to have a legal structure. We informed participants that using the feedback was optional. For most conditions, we used a configuration of the structure-based adaptive policy feedback system that we expected to provide the most usability benefits, as shown in Figure 1.

RQ2: Structure-Based vs. String-Based Adaptiveness *What security and usability impact do structure-based and string-based adaptive policies have relative to each other, as well as relative to a non-adaptive policy?* To analyze this question, we studied only the higher security 3c12 policy in three versions: no adaptiveness; structure-based adaptiveness; and string-based adaptiveness [42]. However, the string-based adaptive policy does not include a feedback system. In contrast to a structure-based system, how to craft a minimally different password securely in a string-based system is non-obvious. The following three conditions address this research question:

- **3c12**: Previously introduced.
- **Struct$_\mathrm{M}$**: Previously introduced.
- **String$_\mathrm{M}$**: 3c12 with 10^5 passwords banned.

RQ3: Varying Blacklist Sizes *What are the usability and security consequences of different size blacklists?* As users create accounts on a structure-based adaptive system, more structures will be banned, potentially making password creation more frustrating for users whose desired structures are banned. As detailed in Section 3.3, we created blacklists of varying sizes to simulate different points in time in the adoption of an adaptive system, as the number of banned structures increases.

To measure the increasing difficulty of creating valid passwords in the presence of larger blacklists, as well as the theoretical security benefits of larger blacklists, we compared the following conditions. All are based on 3c12. Blacklist details are shown in Table 1.

- **3c12**: Previously introduced.
- **Struct$_\mathrm{M}$**: Previously introduced.
- **Struct$_\mathrm{L}$**: 3c12 with 8,940 banned structures (corresponding to banning 1,000,000 passwords).
- **Struct$_\mathrm{XL}$**: 3c12 with 48,199 banned structures (corresponding to banning 10,000,000 passwords).

The remaining research questions involve different modes of feedback to the user when password creation fails.

RQ4: Number of Suggested Modifications *What are the usability and security consequences of presenting the user with more or fewer suggested modifications to their rejected password?* Whereas Struct$_\mathrm{M}$ shows one suggested modification, we also tested a condition that shows (at the same time) three suggested modifications. A participant could choose among the three hints, or choose a completely different password. We also evaluated a condition with hints disabled.

- **Struct$_\mathrm{M}$**: Previously introduced (one hint).
- **Struct$_\mathrm{M}$3Hint**: Identical to Struct$_\mathrm{M}$, but with the feedback system showing three examples of possible modifications to the rejected password that lead to a permitted structure.
- **Struct$_\mathrm{M}$NoHint**: Identical to Struct$_\mathrm{M}$, but with no suggested modifications shown to the user.

RQ5: Insertion vs. Substitution Feedback Our standard method of suggesting modifications to a banned structure would either propose inserting a character or substituting a character. The standard implementation chose among those two possibilities with equal probability. *Does suggesting just insertions or just substitutions affect usability?*

- **Struct$_{\mathrm{M}}$**: Previously introduced.
- **Struct$_{\mathrm{M}}$Ins**: Identical to Struct$_{\mathrm{M}}$, but only offering feedback with suggestions of character insertions.
- **Struct$_{\mathrm{M}}$Sub**: Identical to Struct$_{\mathrm{M}}$, but only offering feedback with suggestions of character substitutions.

RQ6: Preventing Shoulder Surfing of Suggestions We speculate that suggesting modifications to banned structures could potentially aid in shoulder-surfing or screen-scraping attacks because both the rejected password itself and the suggested character to be inserted/substituted are displayed (see Figure 1). *Can the usability of the feedback system be preserved while limiting the potentially sensitive information shown on screen?* To answer this, we compare four conditions with different amounts of potentially sensitive information shown in the feedback interface.

- **Struct$_{\mathrm{M}}$**: Previously introduced.
- **Struct$_{\mathrm{M}}$Hyb**: Like Struct$_{\mathrm{M}}$ in that the rejected password is still shown on screen. However, suggestions instead relate to inserting or substituting a particular class of characters (e.g., a digit) rather than a specific character.
- **Struct$_{\mathrm{M}}$S**: Like Struct$_{\mathrm{M}}$Hyb, except all characters (rejected password and suggested modification) are replaced with a representation of their character class. See Figure 2.
- **Struct$_{\mathrm{M}}$SV**: Identical to Struct$_{\mathrm{M}}$S, except the interface displays the password's structure in real time as the participant types it, rather than only when a password is rejected. The intention was to help users understand the concept of character classes via a real-time example during creation.

3.5 Measuring Password Strength

To evaluate the passwords created in each condition, we analyze general password-composition characteristics, such as average length, inclusion of a variety of character classes, as well as password guessability [49], which models how many guesses a simulated attacker would make to guess a given fraction of a password set. To compute password guessability, we use the Password Guessability Service (PGS) in its recommended configuration (including the cracking methods: Probabilistic Context Free Grammars, Markov Models, Neural Networks, John the Ripper, and Hashcat), which combines several guessing attacks. This approach has previously been shown to be a conservative estimate of an expert in password forensics [49].

Modeling how an attacker would optimally attack a set of passwords created under an adaptive policy raises a number of subtle issues. In a structure-based adaptive policy configured such that a single usage causes a structure to be banned, successfully guessing a password with a particular structure implies that an attacker should avoid making additional guesses with the same structure. Similarly, if an attacker could somehow learn the list of used/banned structures, an attack could be refined by only attempting guesses with those structures. In our tests, we assume that blacklisted structures are unknown to the attacker, and that a structure is banned only after multiple passwords with the same structure are created, making it difficult for an attacker to determine that all passwords with a given structure were guessed and thus benefit from ceasing to make guesses with that structure. With this, we assume that sufficient rate-limiting and/or CAPTCHA solutions are implemented to prevent an attacker from abusing the password creation process to learn details about the adaptive policy's blacklist.

Because an attacker will not know at what point during the adaptive process a particular password was created, they will not be able to exclude potential guesses with particular character-class structures. Thus, we intentionally did *not* modify the computed PGS results to account for different blacklist sizes. Similarly, PCFG results that were used to create the blacklists may still be valid guesses for passwords created early in the adaptive process, yet the attacker does not know which passwords those are.

Hence, our guessability results compare the strength of passwords created earlier during the use of an adaptive policy (e.g., in Struct$_{\mathrm{M}}$) to those created later during the use of an adaptive policy (e.g., in Struct$_{\mathrm{L}}$) to those created without an adaptive policy (e.g., in 3c12).

3.6 Statistical Testing

For our usability metrics, we first performed omnibus statistical tests across all conditions. For omnibus comparisons, we use Kruskal-Wallis (KW) tests for quantitative data and Pearson's Chi-squared tests for categorical data. If the omnibus test was significant, we performed pairwise tests of pre-selected contrasts that correspond with each of our research questions. For pairwise comparisons, we use the Mann-Whitney U tests for quantitative data and Chi-squared tests (Fisher's Exact test when there are small bins) for categorical data. We use non-parametric statistical tests to avoid making assumptions about our data's distribution.

In particular, we made pairwise comparisons between each of the following groups of conditions: varying blacklist sizes (3c12, String$_{\mathrm{M}}$, Struct$_{\mathrm{M}}$, Struct$_{\mathrm{L}}$, Struct$_{\mathrm{XL}}$); the number of hints (Struct$_{\mathrm{M}}$NoHint, Struct$_{\mathrm{M}}$, Struct$_{\mathrm{M}}$3Hint); the type of suggestion (3c12, Struct$_{\mathrm{M}}$Ins, Struct$_{\mathrm{M}}$Sub, Struct$_{\mathrm{M}}$); stopping shoulder-surfing and screen-scraping (Struct$_{\mathrm{M}}$NoHint, Struct$_{\mathrm{M}}$, Struct$_{\mathrm{M}}$SV, Struct$_{\mathrm{M}}$S, Struct$_{\mathrm{M}}$Hyb). In each set, we compared each condition to all other conditions in that group. For all set of pairwise contrasts, we corrected for multiple testing using Holm-Bonferroni correction (HC).

For comparing the results of our simulated cracking attacks, we used a Log-Rank test, a statistical method used in survival analysis [28]. This test compares two guessing curves and takes into account whether a password was guessed, as well as at what point guessing stops. In this way, we can use all the data in our guessing curves for the statistical tests. All statistical tests use a significance level of $\alpha = .05$.

3.7 Limitations

Our methodology, which is similar to that employed by prior password research [30, 32, 45, 47, 48], has a number of limitations. By testing password recall once after a few minutes and once again after a few days, our study investigated password use that lies in between frequent and rare use. As such, we are not able to make strong statements about participants' ability to remember passwords in our study over long periods of time.

Condition	# Participants	Length	Uppercase	Lowercase	Digits	Symbols	Guessed (%)	Create difficult (%)	Create attempts	Create time (s)	Create confusing (%)	Recall attempts	Recall time (s)
3c12	163	13.9	1.6	7.9	3.1	1.3	49.1	28.2	1.50	47.9	9.82	1.80	26.2
String_M	216	13.6	1.7	7.4	3.2	1.3	46.8	26.4	1.50	48.5	6.94	1.82	23.3
Struct_M	213	14.1	1.8	7.5	3.5	1.3	25.4	31.5	2.08	50.4	9.39	1.57	25.3
Struct_L	159	14.1	1.6	7.3	3.8	1.4	17.6	43.4	2.51	48.4	19.5	1.86	26.6
Struct_XL	247	14.0	1.8	7.2	3.6	1.4	14.1	36.8	2.58	50.8	21.5	1.64	30.3
Struct_M3Hint	216	14.2	2.0	7.6	3.3	1.3	25.0	32.4	1.96	49.2	12.0	1.73	26.5
Struct_MNoHint	163	13.9	2.0	7.3	3.3	1.3	19.6	40.5	2.12	48.0	17.8	1.67	24.9
Struct_MIns	207	14.4	1.9	7.9	3.3	1.4	23.1	32.9	1.97	54.3	13.5	1.78	27.0
Struct_MSub	202	14.0	1.8	7.7	3.2	1.3	23.3	37.1	2.06	56.0	11.4	1.76	33.6
Struct_MHyb	206	14.0	2.0	7.6	3.1	1.2	29.1	45.1	2.17	51.7	14.1	1.68	30.1
Struct_MS	209	13.8	2.1	7.0	3.6	1.3	31.1	33.0	2.00	54.3	12.9	1.90	28.2
Struct_MSV	204	14.0	1.9	6.9	3.7	1.5	25.5	32.4	1.88	58.5	14.7	1.98	28.3
Struct_M4c8	214	11.1	1.6	5.4	2.8	1.3	60.3	26.2	2.39	42.5	12.6	1.84	23.7

Table 2: Properties of passwords and study measurements, by condition. The second column shows participants who finished part two within three days.

Across our conditions, a relatively high number of participants did not return for part two. We excluded them from our analyses, except for analyzing the dropout rate as an indicator of dissatisfaction with a condition. Users who drop out of a study may behave differently than those who do not, potentially biasing our results.

All of the passwords in our study were collected for this study and were not used to protect real accounts, limiting ecological validity. In contrast to real-world, high-value passwords, study participants would not suffer consequences if they chose a weak password or forgot their password, nor were they incentivized to adopt their normal password behavior beyond our request that they do so. Two recent studies investigated the degree to which passwords collected for research studies resemble real, high-value accounts, and both concluded that passwords created during studies can resemble real, high-value passwords, yet are not a perfect proxy [18, 37].

While password-guessing approaches are most successful at modeling passwords given closely matched training data [30, 36], no major leaks of passwords contain passwords created under 3c12 and 4c8 policies. To compensate, we trained a probabilistic context-free grammar on the subset of passwords from large-scale leaks that fit those policies. We also used this grammar to model large numbers of likely passwords to create the blacklists. While having very large sets of real 3c12 and 4c8 passwords would have been strictly more accurate, no such sets are currently available to researchers.

For the reasons described in Section 3.5, we believe PGS models a reasonable attacker even for adaptive policies. Conceivably, however, some other strategy for ordering guesses against adaptive policies could prove to be more effective. That said, we are not currently aware of any such attack.

4. RESULTS

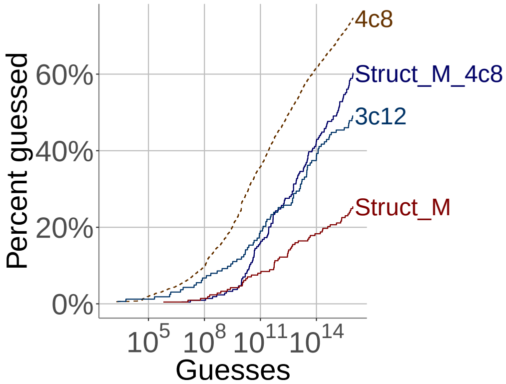
We find that an adaptive policy with a large blacklist dramatically increased the security of passwords. Surprisingly, this large increase in security is accompanied by only a small impact on usability. We tested numerous interface modifications to mitigate the decrease we expected in usability. In the absence of substantial usability decreases, however, these interface modifications have minimal impact on either security or usability. We detail general password characteristics by condition in Table 2; guessability in Figure 3; and usability in Table 3.

Participants received 55 cents for the first part of our study and 70 cents for the second. Of the 3,391 participants who began our study, 2,619 finished part one, 1,975 returned for part two within three days of receiving our invitation to return, and 1,799 finished part two of the study within three days of receiving that invitation. Other than the discussion of dropout rates, our analysis focuses only on the 1,799 participants who finished the entire study. Participants for whom we detect electronic copy-pasting from keystroke timing data almost without exception report that they wrote down their password in the survey, which suggests that participants truthfully disclosed rates of password storage. The number of participants per condition is shown in Table 2. 53% of participants reported being male, 46% female, and the remaining 1% declined to answer. Participants' mean age was 29 years (median 29).

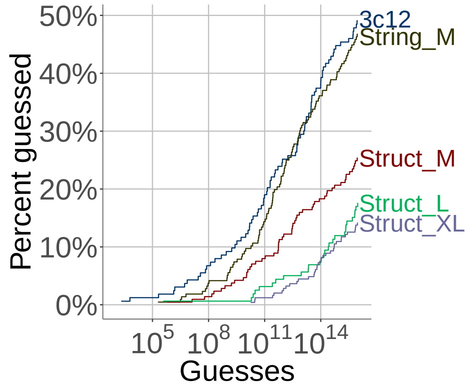
4.1 Impact of Structure-Based Adaptation

To examine the effect of implementing an adaptive policy, we compared 3c12 to Struct_M and 4c8 to Struct_M4c8. These two pairs each compare a password-composition policy with a structure-based adaptive blacklist to one without.

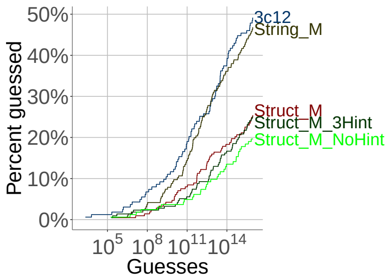
The inclusion of structure-based blacklists had a profound effect on security for both the 3c12 and 4c8 policies. As shown in Figure 3a, after 10^{16} guesses, PGS had correctly guessed roughly half as many passwords in Struct_M and Struct_M4c8 (with the adaptive policy) compared to 3c12 and 4c8 (without the adaptive policy), respectively. The difference between 3c12 and Struct_M is statistically significant (Log-Rank test, $X^2(1) = 23.9$, $p < 0.001$). Because the data for 4c8 was collected for a prior study, we did not perform statistical testing on that comparison.



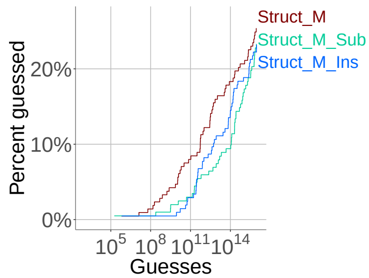
(a) Traditional policies and adaptive policies.



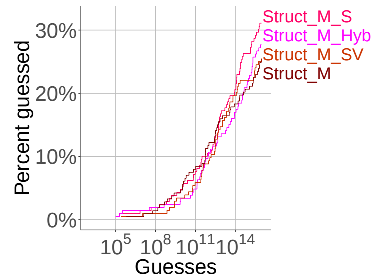
(b) Structure-based adaptive policy conditions with different blacklist sizes.



(c) Varying number of hints for the structure-based adaptive policy conditions.



(d) Varying suggestion type with structure-based adaptive policy.



(e) Conditions with interface changes to prevent shoulder-surfing.

Figure 3: The guessability of each password set. The x-axis shows the guess number (logarithmic scale). The y-axis shows the percent guessed at that guess number. Lines that are lower represent passwords that are more resistant against guessing attacks.

Along with our hypothesis that adaptive policies would result in more secure passwords, which was supported by our data, we also hypothesized that adaptive policies would result in decreased usability. Surprisingly, the structure-based approaches with medium-sized structure blacklists in Struct$_{\mathrm{M}}$ and Struct$_{\mathrm{M}}$4c8 had only minimal impact on usability over 3c12 and 4c8.

We found no significant differences in our omnibus comparisons across all 12 conditions for these usability metrics. For instance, we did not find the inclusion of an adaptive policy to cause participants to perceive password creation as significantly more difficult or confusing. Similarly, we did not find the inclusion of an adaptive policy to make participants significantly more likely to store their passwords on paper or electronically. Nor did we find the inclusion of an adaptive policy to significantly impact the proportion of participants who were able to recall their password or how many attempts it took them to do so.

The only usability decrease that resulted from the structure-based adaptive policy with a medium-sized blacklist was increasing the number of attempts required to create a password. Specifically, participants required significantly more attempts to create compliant Struct$_{\mathrm{M}}$ passwords than 3c12 passwords (KW, H(1) = 22.9, $p < 0.001$) and Struct$_{\mathrm{M}}$4c8 (KW, H(1) = 46.8, $p < 0.001$). That adaptive policies cause users to require more attempts to create a compliant password is unsurprising, though. By design, adaptive policies must reject candidate passwords to have any effect. Between the adaptive policies, we again observed differences in the number of attempts participants required to create a compliant password. For participants to create a compliant password, Struct$_{\mathrm{M}}$4c8 required significantly more attempts than Struct$_{\mathrm{M}}$ (KW, H(1) = 5.48, $p < 0.019$). Struct$_{\mathrm{M}}$4c8 also required significantly more time to submit a first attempt, whether compliant or not, than Struct$_{\mathrm{M}}$ (KW, H(1) = 7.34, $p < 0.020$).

4.2 Structure-Based vs. String-Based

Our experimental design allows for some limited comparison between the Struct$_{\mathrm{M}}$ and String$_{\mathrm{M}}$ approaches. Because optimal configurations for the Leininger et al. [33] and Schecter et al. [42] approaches have not yet been established, results of these comparisons should not be generalized beyond our particular configurations.

Under the configurations we tested in Struct$_{\mathrm{M}}$ and String$_{\mathrm{M}}$, whose blacklists were built using the same source passwords, we find that Struct$_{\mathrm{M}}$ produces passwords roughly twice as difficult to guess as String$_{\mathrm{M}}$ (46.8% vs 25.4% cracked at cutoff, LogRank, $X^2(1) = 20.9$, p < 0.001), with similar usability results. In fact, the guessability of the string-based String$_{\mathrm{M}}$ did not differ significantly from 3c12 (49.1% vs 46.8% cracked at cutoff, LogRank, $X^2(1) = 0.431$, p $<$ 0.735), which did not have an adaptive component.

Intuitively, the security improvement occurs because blacklisting a structure eliminates many potentially common passwords at once, whereas blacklisting a string eliminates only one. In terms of usability, a key factor is that it is trivial to quickly and automatically suggest a modified password

Agree password creation confusing *Omni.* χ^2_{13}=43.3, p<.001					
cond.1	%	cond.2	%	χ^2_1	p-value
String$_M$	6.94	Struct$_L$	19.5	12.3	.003
		Struct$_{XL}$	21.5	18.2	<.001
Struct$_M$	9.39	Struct$_L$	19.5	7.03	.040
		Struct$_{XL}$	21.5	11.6	.005

Agree password creation difficult *Omni.* χ^2_{13}=39.1, p<.001					
cond.1	%	cond.2	%	χ^2_1	p-value
String$_M$	26.4	Struct$_L$	43.4	11.1	.009
Struct$_M$	31.5	Struct$_M$Hyb	45.1	7.75	.032

Password creation attempts *Omni.* KW χ^2_{13}=143, p<.001					
cond.1	mean	cond.2	mean	χ^2_1	p-value
3c12	1.50	Struct$_M$	2.08	22.9	<.001
		Struct$_M$4c8	2.39	46.8	<.001
String$_M$	1.50	Struct$_L$	2.51	33.3	<.001
		Struct$_{XL}$	2.58	68.0	<.001
Struct$_M$	2.08	3c12	1.50	22.9	<.001
		Struct$_M$4c8	2.39	5.48	.019
		Struct$_{XL}$	2.58	11.2	.004
Struct$_M$SV	1.88	Struct$_M$Hyb	2.17	6.96	.05

Password entry time during creation (s) *Omni.* KW χ^2_{13} =33.9, p=.001					
cond.1	median	cond.2	median	χ^2_1	p-value
Struct$_M$	50.4	Struct$_M$4c8	42.5	7.34	.020

Password entry time during recall (s) *Omni.* χ^2_{13}=37.8, p<.001					
cond.1	median	cond.2	median	χ^2_1	p-value
Struct$_{XL}$	30.3	String$_M$	23.3	11.5	.006

% Cracked (Log-Rank test)					
cond.1	%	cond.2	%	χ^2_1	p-value
3c12	49.1	Struct$_M$	25.4	23.9	<.001
		Struct$_L$	17.6	42.7	<.001
		Struct$_{XL}$	14.1	73.3	<.001
		Struct$_M$3Hint	25.0	23.9	<.001
		Struct$_M$NoHint	19.6	27.9	<.001
String$_M$	46.8	Struct$_L$	17.6	38.9	<.001
		Struct$_M$	25.4	20.9	<.001
		Struct$_{XL}$	14.1	68.6	<.001
Struct$_M$	25.4	Struct$_M$4c8	28.5	16.2	<.001
		Struct$_{XL}$	14.1	11.2	<.003

Table 3: The statistically significant pairwise differences among our metrics.

that is close to the user's original attempt but still guaranteed to pass the structure check. Because any string-based password that is rejected is itself already a popular password, how one might automatically generate a minimally different, yet secure, password is non-obvious.

Although passwords created under Struct$_M$ were significantly more secure than those created under String$_M$, we did not observe significant differences between these two conditions for any of our usability metrics. As we describe later, however, we did find String$_M$ to have significant usability advantages over the structure-based policies configured with larger blacklists.

While more research comparing these approaches is necessary, our results suggest that a system administrator with access to a limited list of passwords with which to generate an initial blacklist should use a structure-based, rather than string-based, approach.

4.3 Varying Blacklist Sizes

Having found that implementing an adaptive system led to far more secure passwords while incurring minimal usability cost, we also explored how varying the size of the blacklists would impact security and usability. As we detailed in Section 3.4, these blacklists of different sizes should primarily be interpreted as proxies for different points in time during the life cycle of an adaptive policy, rather than configuration options. We also evaluated how these structure-based blacklists of different sizes compared to the medium size string-based blacklist. To do so, we compare the following four conditions: 3c12, String$_M$, Struct$_M$, Struct$_L$, Struct$_{XL}$.

The security of the passwords generally increased with the size of the blacklist, as shown in Figure 3b. Compared to 3c12, significantly fewer Struct$_L$ and Struct$_{XL}$ passwords were guessed (Log-Rank test, 3c12 vs Struct$_L$, $X^2(1) = 42.7$, $p < 0.001$; 3c12 vs Struct$_{XL}$, $X^2 = 73.3$, $p < 0.001$). Surprisingly, the guessability of Struct$_L$ and Struct$_{XL}$ did not differ significantly, suggesting that at structure blacklists of those sizes, the probability of a user creating a password with the next most common structure over any other permitted structure is very small.

Unsurprisingly, password creation generally required less effort in conditions with smaller blacklists. In essence, password creation becomes harder over time in an adaptive system. Struct$_M$ required significantly fewer creation attempts than Struct$_{XL}$ (KW, H(1) = 11.2, $p < 0.004$). In contrast, the time to create passwords on the first attempt did not differ significantly across conditions. This finding makes sense because participants in all conditions were shown the same text and interface during the first creation attempt.

Participants in conditions with smaller blacklists found password creation less difficult than those in conditions with larger blacklists. Participants in Struct$_M$ rated password creation as less confusing than participants in Struct$_L$ (Chi-squared, $X^2(1) = 7.03$, $p < 0.040$) or in Struct$_{XL}$ (Chi-squared, $X^2(1) = 11.6$ $p < 0.004$).

Despite these differences during password creation, we observed few differences across conditions in terms of password recall, suggesting that password memorability does not decrease significantly for users who create passwords later in the adaptive process. More precisely, the rate at which participants stored their passwords did not differ significantly across conditions (omnibus $X^2(13) = 16.3$, $p = 0.233$). The number of attempts participants required to recall their password also did not differ significantly across conditions (omnibus KW, H(13) = 9.21, $p = 0.757$).

All structure-based blacklists we tested resulted in more secure passwords than the string-based blacklist we tested.

That is, compared to String_{M}, fewer Struct_{M}, Struct_{L}, and $\text{Struct}_{\text{XL}}$ passwords were guessed (Log-Rank test, String_{M} vs Struct_{M}, $X^2(1) = 20.9$, $p < 0.001$; String_{M} vs Struct_{L}, $X^2(1) = 38.9$, $p < 0.001$; String_{M} vs $\text{Struct}_{\text{XL}}$, $X^2(1) = 68.6$, $p < 0.001$).

Although it was less secure than the structure-based adaptive conditions, the string-based adaptive condition was generally more usable, requiring significantly fewer creation attempts than condition Struct_{L} (KW, H(1) = 44.0, $p < 0.001$) or $\text{Struct}_{\text{XL}}$ (KW, H(1) = 68.0, $p < 0.001$). Participants in String_{M} rated password creation as significantly less difficult (Chi-squared, $X^2(1) = 11.1$, $p < 0.008$) and less confusing (Chi-squared, $X^2(1) = 12.3$, $p < 0.003$) than participants in Struct_{L}. Participants in String_{M} required less time to recall passwords than $\text{Struct}_{\text{XL}}$ participants (Chi-squared, $X^2(1) = 11.5$), $p = 0.006$) even though, as stated earlier, we did not observe significant differences in the memorability of those passwords.

4.4 Number of Suggested Modifications

We initially hypothesized that structure-based adaptive policies would cause a profound loss in usability. Therefore, we focused a number of conditions on the feedback given to users when their password was rejected. In those cases, the system would suggest modifications to the user's rejected password to make it compliant.

We tried varying the number of suggested modifications (one suggested modification versus three), as well as not suggesting any modifications. However, varying the number of suggested modifications did not have an impact on either security or usability.

In particular, we made pairwise comparisons across conditions Struct_{M}NoHint, Struct_{M}, and Struct_{M}3Hint. We did not observe statistically significant differences in the relative guessability of any of the following three condition pairs (Log-Rank test, Struct_{M}3Hint vs. Struct_{M} $X^2(1) = 0.06$, p = 1.0; Struct_{M}NoHint vs. Struct_{M} $X^2(1) = 1.887$, p = 0.678; Struct_{M}NoHint vs. Struct_{M}3Hint $X^2(1) = 1.415$, p = 0.703). Similarly, we did not find any pairwise comparisons for our usability metrics to have statistically significant differences.

We also calculated how many participants saw a hint, as well as how many accepted the hint's advice. In condition Struct_{M}, 75 of 213 participants saw at least one hint generated by the adaptive password policy during password creation, similar to the hint shown in Figure 1. Of those 75, slightly less than half (34) did not accept the advice shown in the hint and attempted to create an entirely new password, while the remaining 41 participants followed the guidance provided by the feedback.

4.5 Insertion vs. Substitution Feedback

We also examined the type of suggestions the adaptive system makes for rejected passwords. We compared Struct_{M}Ins, Struct_{M}Sub, and Struct_{M}, which respectively gave participants feedback that suggested either inserting a character, substituting a character, or one of the two (with equal probability). The locations of the character insertion/substitution suggestions were chosen randomly. We did not observe any significant differences in usability across these conditions.

4.6 Shoulder Surfing of Suggestions

As detailed in Section 3.4, we varied the suggested modifications in ways designed to minimize the information shown on screen, experimenting with showing structures instead of the actual password in either the suggestions, and as the user types their password. We expected that minimizing this information would decrease usability, yet would minimize the advantage to a shoulder-surfing adversary. To evalute this, we compared Struct_{M}, Struct_{M}Hyb, Struct_{M}S, and Struct_{M}SV. Because the extra information gleaned from shoulder surfing is not modeled in our guessability analyses, we did not expect to observe differences in guessability.

As expected, these conditions did not differ significantly in guessability. Surprisingly, though, we also observed minimal impact on usability. Although 45% of participants in Struct_{M}Hyb said creating a password was difficult, which was marginally higher than the proportion of participants Struct_{M} (32%) who shared the same sentiment (Chi-squared, $X^2(1) = 7.75$, $p < 0.032$), we did not observe any other significant differences in usability.

5. DISCUSSION

Overall, we found that applying a structure-based adaptive policy to 3c12 was beneficial, substantially increasing security with a comparatively mild negative effect on usability. The effect on security was dramatic; about half as many passwords were cracked in condition Struct_{M} as in 3c12. Surprisingly, although participants on average required more attempts to create compliant passwords in condition Struct_{M} than 3c12, participants did not rate password creation as significantly more difficult. More importantly, the number of attempts required to recall their password, password entry time, and the fraction of participants who stored their password did not differ between conditions, suggesting that the structure-based adaptive policy does not negatively affect password memorability.

Varying the structure blacklist size, our proxy for an increase in the number of users of a given adaptive system, had profound effects on the security of passwords. As more users join the system and more structures are banned, new users are creating far more secure passwords than the initial users of the system. As expected, larger blacklists caused participants to require more creation attempts, yet this mostly did not increase participants' perceived difficulty of the task, in contrast to prior experiments (e.g., [45]). Interestingly, $\text{Struct}_{\text{XL}}$ had no security benefits over Struct_{L}, suggesting that the security benefits may have diminishing returns as the structure blacklist grows. Taking into account these diminishing returns, as well as the security disadvantages of blacklisting a structure after a single use (Section 3.5), we recommend blacklisting a structure only after multiple uses. Based on the diminishing returns of blacklisting structures, it could be beneficial to increase the number of uses before a structure is blacklisted as the number of passwords in the system increases. For systems with huge user bases (e.g. Google, Facebook, Twitter) this concept may become more important. We also suggest bootstrapping this system with the few thousand most common structures and letting the blacklist grow over time; this significantly increased resistance to guessing attacks with minimal usability sacrifices.

Neither varying the number nor removing hints altogether had a significant impact. Only about half of participants who saw a hint (34 of 75) in condition Struct_{M} used the suggested password. This could be because participants felt they could make passwords that were more memorable, yet would still satisfy the requirements, or felt it would be more secure to use their own changes.

Based on prior work [20], we expected participants to find insertion suggestions more usable than substitution suggestions. However, we did not find this to be the case. At the same time, we found no significant differences with respect to the resistance of such passwords to guessing attacks.

A drawback of any password-creation feedback interface is that it could risk revealing information to attackers about the password through shoulder-surfing attacks. We hoped to minimize the impact of shoulder surfing by providing somewhat obfuscated feedback to participants. With minor exceptions, we found no significant differences according to our strength and usability metrics. As a result, we recommend the techniques used in Struct_{M}S, or Struct_{M}SV if real-time feedback is desired.

6. CONCLUSION

We evaluated string- and structure-based adaptive password policies, finding that adaptive policies provide significant security benefit with seemingly little usability cost, and should be considered for use in environments with large numbers of users. To balance usability and security, we recommend augmenting a strong password-composition policy with a structure-based adaptive system.

Surprisingly, the feedback system we tested did not improve usability as we had expected. Regardless of the type of feedback provided, participants made significantly stronger passwords with structure-based blacklists than without them, leading us to speculate that simply instructing participants who attempted to create blacklisted passwords to try to create a password with an uncommon sequence of character classes was sufficient; this should be investigated in future work. We find that obfuscating suggested passwords by their character-class representations, or not giving feedback at all, to be as usable as feedback approaches that are more vulnerable to shoulder surfing.

7. ACKNOWLEDGMENTS

This research was supported in part by PNC Financial Services Group, and by a grant from NATO through Carnegie Mellon CyLab. The authors acknowledge KoreLogic for technical assistance with PathWell.

8. REFERENCES

[1] Abdou, A., Barrera, D., and van Oorschot, P. What lies beneath? Analyzing automated SSH bruteforce attacks. In *Proc. Passwords* (2015).

[2] Barrett, B. 7 password experts on how to lock down your online security. *Wired*, 2016. https://www.wired.com/2016/05/password-tips-experts/.

[3] Bloom, B. H. Space/time trade-offs in hash coding with allowable errors. *Commun. ACM 13*, 7 (July 1970), 422–426.

[4] Bonneau, J. The Gawker hack: How a million passwords were lost, 2010. https://www.lightbluetouchpaper.org/2010/12/15/the-gawker-hack-how-a-million-passwords-were-lost/.

[5] Bonneau, J. *Guessing human-chosen secrets*. PhD thesis, University of Cambridge, 2012.

[6] Bonneau, J., and Schechter, S. Towards reliable storage of 56-bit secrets in human memory. In *Proc. USENIX Security* (2014).

[7] Burr, W. E., Dodson, D. F., Newton, E. M., Perlner, R. A., Polk, W. T., Gupta, S., and Nabbus, E. A. Electronic authentication guideline. Tech. rep., NIST, 2011.

[8] Castelluccia, C., Durmuth, M., and Perito, D. Adaptive password-strength meters from Markov models. In *Proc. NDSS* (2012).

[9] Choong, Y.-Y., Theofanos, M., and Hung-Kung, L. United States Federal Employees' Password Management Behaviors. Tech. rep., NIST, 2014.

[10] Chou, H.-C., Lee, H.-C., Yu, H.-J., Lai, F.-P., Huang, K.-H., and Hsueh, C.-W. Password cracking based on learned patterns from disclosed passwords. *IJICIC* (2013).

[11] Das, A., Bonneau, J., Caesar, M., Borisov, N., and Wang, X. The Tangled Web of Password Reuse. In *Proc. NDSS* (2014).

[12] Davies, C. Millions of eHarmony passwords leaked. *Slash Gear*, 2012. https://www.slashgear.com/millions-of-eharmony-passwords-leaked-07232691/.

[13] Dell'Amico, M., Michiardi, P., and Roudier, Y. Password strength: An empirical analysis. In *Proc. INFOCOM* (2010).

[14] Duckett, C. Login duplication allows 20m Alibaba accounts to be attacked. *ZDNet*, 2016. http://www.zdnet.com/article/login-duplication-allows-20m-alibaba-accounts-to-be-attacked/.

[15] Dürmuth, M., Angelstorf, F., Castelluccia, C., Perito, D., and Chaabane, A. OMEN: Faster password guessing using an ordered markov enumerator. In *ESSoS*. 2015.

[16] Dürmuth, M., Chaabane, A., Perito, D., and Castelluccia, C. When privacy meets security: Leveraging personal information for password cracking. *CoRR* (2013).

[17] E2 Solutions. CW Government Travel, 2016.

[18] Fahl, S., Harbach, M., Acar, Y., and Smith, M. On the ecological validity of a password study. In *Proc. SOUPS* (2013).

[19] Florêncio, D., and Herley, C. Where do security policies come from? In *Proc. SOUPS* (2010).

[20] Forget, A., Chiasson, S., van Oorschot, P. C., and Biddle, R. Improving text passwords through persuasion. In *Proc. SOUPS* (2008).

[21] Geuss, M. Mozilla: Data stolen from hacked bug database was used to attack Firefox. *Ars Technica*, 2015. https://arstechnica.com/security/2015/09/mozilla-data-stolen-from-hacked-bug-database-was-used-to-attack-firefox/.

[22] Goodin, D. Hackers expose 453,000 credentials allegedly taken from Yahoo service. *Ars Technica*, 2012. http://arstechnica.com/security/2012/07/yahoo-service-hacked/.

[23] Goodin, D. Why passwords have never been weaker-and crackers have never been stronger. *Ars*

Technica, 2012. https://arstechnica.com/security/2012/08/passwords-under-assault/.

[24] Goodin, D. “thereisnofatebutwhatwemake”-turbo-charged cracking comes to long passwords. *Ars Technica*, 2013. https://arstechnica.com/security/2013/08/thereisnofatebutwhatwemake-turbo-charged-cracking-comes-to-long-passwords/.

[25] Goodin, D. Once seen as bulletproof, 11+ million Ashley Madison passwords already cracked. *Ars Technica*, 2015. https://arstechnica.com/security/2015/09/once-seen-as-bulletproof-11-million-ashley-madison-passwords-already-cracked/.

[26] Graham, R. Notes on the Ashley-Madison dump, 2015. http://blog.erratasec.com/2015/08/notes-on-ashley-madison-dump.html.

[27] Habib, H., Colnago, J., Melicher, W., Ur, B., Segreti, S., Bauer, L., Christin, N., and Cranor, L. F. Password creation in the presence of blacklists. In *Proc. USEC* (2017).

[28] Harrington, D. P., and Fleming, T. R. A class of rank test procedures for censored survival data. *Biometrika 69*, 3 (1982), 553–566.

[29] Juels, A., and Rivest, R. L. Honeywords: Making password-cracking detectable. In *Proc. CCS* (2013).

[30] Kelley, P. G., Komanduri, S., Mazurek, M. L., Shay, R., Vidas, T., Bauer, L., Christin, N., Cranor, L. F., and Lopez, J. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In *Proc. IEEE SP* (2012).

[31] Komanduri, S. *Modeling the adversary to evaluate password strengh with limited samples*. PhD thesis, CMU, 2015.

[32] Komanduri, S., Shay, R., Kelley, P. G., Mazurek, M. L., Bauer, L., Christin, N., Cranor, L. F., and Egelman, S. Of passwords and people: measuring the effect of password-composition policies. In *Proc. CHI* (2011).

[33] KoreLogic. libpathwell: Pam module and library for auditing/enforcing password topology histogram wear-leveling, 2015. https://github.com/KoreLogicSecurity/libpathwell.

[34] Leininger, H. LibPathWell 0.6.1 Released, 2015. https://blog.korelogic.com/blog/2015/07/31/libpathwell-0_6_1.

[35] Love, D. Apple on iCloud breach: It’s not our fault hackers guessed celebrity passwords. *International Business Times*, 2014. http://www.ibtimes.com/apple-icloud-breach-its-not-our-fault-hackers-guessed-celebrity-passwords-1676268.

[36] Ma, J., Yang, W., Luo, M., and Li, N. A study of probabilistic password models. In *Proc. IEEE SP* (2014).

[37] Mazurek, M. L., Komanduri, S., Vidas, T., Bauer, L., Christin, N., Cranor, L. F., Kelley, P. G., Shay, R., and Ur, B. Measuring password guessability for an entire university. In *Proc. CCS* (2013).

[38] Microsoft. Microsoft SMB Protocol and CIFS Protocol Overview.

[39] Narayanan, A., and Shmatikov, V. Fast dictionary attacks on passwords using time-space tradeoff. In *Proc. CCS* (2005).

[40] Rao, A., Jha, B., and Kini, G. Effect of grammar on security of long passwords. In *Proc. CODASPY* (2013).

[41] Redman, R. PathWell Topologies, 2014. https://blog.korelogic.com/blog/2014/04/04/pathwell_topologies.

[42] Schechter, S., Herley, C., and Mitzenmacher, M. Popularity is everything: A new approach to protecting passwords from statistical-guessing attacks. In *Proc. HotSec* (2010).

[43] Schneier, B. Myspace passwords aren’t so dumb. *Wired*, 2006. http://archive.wired.com/politics/security/commentary/securitymatters/2006/12/72300.

[44] Shay, R., Kelley, P. G., Komanduri, S., Mazurek, M. L., Ur, B., Vidas, T., Bauer, L., Christin, N., and Cranor, L. F. Correct horse battery staple: Exploring the usability of system-assigned passphrases. In *Proc. SOUPS* (2012).

[45] Shay, R., Komanduri, S., Durity, A. L., Huh), P. S., Mazurek, M. L., Segreti, S. M., Ur, B., Bauer, L., Christin, N., and Cranor, L. F. Can long passwords be secure and usable? In *Proc. CHI* (2014).

[46] Trustwave. 2014 global security report. https://www.trustwave.com/Resources/Library/Documents/2014-Trustwave-Global-Security-Report.

[47] Ur, B., Alfieri, F., Aung, M., Bauer, L., Christin, N., Colnago, J., Cranor, L. F., Dixon, H., Naeini, P. E., Habib, H., Johnson, N., and Melicher, W. Design and evaluation of a data-driven password meter. In *Proc. CHI* (2017).

[48] Ur, B., Kelley, P. G., Komanduri, S., Lee, J., Maass, M., Mazurek, M., Passaro, T., Shay, R., Vidas, T., Bauer, L., Christin, N., and Cranor, L. F. How does your password measure up? The effect of strength meters on password creation. In *Proc. USENIX Security* (2012).

[49] Ur, B., Segreti, S. M., Bauer, L., Christin, N., Cranor, L. F., Komanduri, S., Kurilova, D., Mazurek, M. L., Melicher, W., and Shay, R. Measuring real-world accuracies and biases in modeling password guessability. In *Proc. USENIX Security* (2015).

[50] Vance, A. If your password is 123456, just make it HackMe. *The New York Times*, 2010. http://www.nytimes.com/2010/01/21/technology/21password.html?mcubz=1.

[51] Veras, R., Collins, C., and Thorpe, J. On the semantic patterns of passwords and their security impact. In *Proc. NDSS* (2014).

[52] Weir, M., Aggarwal, S., Collins, M., and Stern, H. Testing metrics for password creation policies by attacking large sets of revealed passwords. In *Proc. CCS* (2010).

[53] Weir, M., Aggarwal, S., de Medeiros, B., and Glodek, B. Password cracking using probabilistic context-free grammars. In *Proc. IEEE SP* (2009).

A Second Look at Password Composition Policies in the Wild: Comparing Samples from 2010 and 2016

Peter Mayer[⋆], Jan Kirchner[†], Melanie Volkamer[⋆‡]
[⋆] SECUSO - Security, Usability, Society, Technische Universität Darmstadt
[†] Institute of Psychology, Technische Universität Darmstadt
[‡] Privacy and Security Research Group, Karlstad University
{peter.mayer,melanie.volkamer}@secuso.org

ABSTRACT

In this paper we present a replication and extension of the study performed by Florêncio and Herley published at SOUPS 2010. They investigated a sample of US websites, examining different website features' effects on the strength of the website's password composition policy (PCP). Using the same methodology as in the original study, we re-investigated the same US websites to identify differences over time. We then extended the initial study by investigating a corresponding sample of German websites in order to identify differences across countries. Our findings indicate that while the website features mostly retain their predicting power for the US sample, only one feature affecting PCP strength translates to the German sample: whether users can choose among multiple alternative websites providing the same service. Moreover, German websites generally use weaker PCPs and, in particular, PCPs of German banking websites stand out for having generally low strength PCPs.

1. INTRODUCTION

Creating a password usually requires adherence to a password composition policy (PCP). Florêncio and Herley [5] analysed the PCPs of 75 different websites in 2010 and reported a high diversity. They investigated several website features (e.g. whether the user name is publicly visible, the value of the resources protected by the password, or whether the website advertises on other websites) in order to isolate those features that influence the PCPs' strength. They found that the security-related features of a website did not correlate with the PCPs' strength. Instead, those websites which were not affected by the consequences of bad usability (e.g. government sites, because users have no alternative), had the strongest PCPs.

However, it has been several years since their investigation and one might wonder: Has the landscape of PCPs on the Internet changed since their initial investigation? Have Internet PCPs become more or less strict? Have the originally-analysed features lost or gained influence on the strength of PCPs? Also, the original study only examined US websites. Thus, it remains an open question whether the features have the same influence on PCP strength of other countries' websites.

We decided to investigate these questions in a replication of Florêncio and Herley's study [5] (original study), extended by the inclusion of a corresponding sample of German websites. Thereby, our goal was not only to revisit the original research questions of Florêncio and Herley [5], but to explore whether a comparison of PCPs over time, and across country borders, yields new findings.

Our results indicate that the US PCPs have become, on average, stronger in the intervening years. For the US sample, most features retain the predictive power with respect to the strength of a website's PCP found in the original study. However, only one of the features used in the original study emerges as a reliable predictor for the German sample: websites facing a potential loss of users due to poor usability are more likely to have weaker PCPs. Furthermore, German websites employ, on average, weaker PCPs than US websites (see Figure 1). In particular, the PCPs used by German banking websites are significantly weaker than those of US banking websites and also exhibit the lowest average PCP strength in the German sample. This finding, combined with the identified predictive factor, "user has choice" may indicate that German banks are especially keen

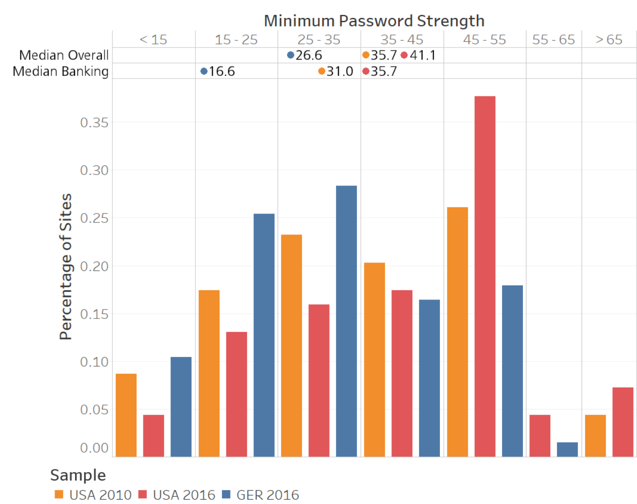


Figure 1: Histogram of PCP strengths (according to the method used in [5]) of the three samples.

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

to offer the most user-friendly experience in order to not lose customers due to poor usability.

In the following, we first describe the original study by Florêncio and Herley in detail and present related work. Then, we describe the methodology of this replication study, including our research questions. Thereafter, we present the results for the individual samples in relation to our four research questions. The results are discussed with respect to our research questions afterwards. Finally, we present the conclusions, which we draw from this study.

2. ORIGINAL STUDY

In the original study, Florêncio and Herley [5] investigated the strength of PCPs of 75 different US websites. They sampled websites according to different categories: 15 top traffic websites (determined by Quantcast[1] traffic rank of 20 or higher), 8 high traffic websites (determined by Quantcast traffic rank of 101 to 110), 8 medium traffic sites (determined by Quantcast traffic rank 1001 to 1010), 9 banking websites (the top ranked US banks and brokerages), 10 websites of large universities (determined by their 2006 enrolment numbers), 10 websites of the universities with the top computer science departments (determined as per U.S.News), the 10 government websites with the highest traffic (determined by Quantcast traffic rank), and 7 miscellaneous sites “for comparison interest” [5]. Websites without account systems were simply dropped from their sample. Two websites appeared in two different categories each and were considered in both categories during the analysis.

For all these websites, the PCPs were determined using the following procedure. If possible, an account was created on the website. In case this was not possible, the authors relied on published password policies. To find them, they performed a web search and considered only the first PCP they found.

For each of the identified PCPs, they calculated the minimum strength of the PCP using $N_{min} \cdot \log_2(C_{min})$ where N_{min} is the minimum length allowed by the PCP and C_{min} is the cardinality of the minimum character set required to fulfil the PCP[2]. Their reasoning behind choosing a minimum strength measure is that the intent of PCPs is to enforce a minimum strength among passwords on a website and users usually wont choose passwords which are much more secure. They acknowledge that it is not a perfect measure and does not model guessing resistance, but argue that it preserves the ordering of PCPs in terms of burden on the user. In this paper, we refer to the minimum strength of a PCP without explicitly including the term *minimum* for readability's sake. We simply refer to it as *PCP strength*.

Using this sample of PCPs and their respective strengths, they investigate the effects of the following website features:

1. Observation and evidence with regard to breaches.
2. The size of the service as determined by traffic rank and number of users.
3. Whether the user name is public on the website.
4. The value of the resources protected based on the type of website.
5. The extractable value of the resources protected based on the monetisation of data gained from breaches.
6. Who lives with the consequences of a breach as determined by the policies of the websites.
7. Whether third party advertising is accepted on that website as determined by the Quantcast advertising information.
8. Whether the site advertises as determined by the use of Google Sponsored Links of that website.
9. Whether the user can choose alternative websites offering the same service.

Thereby, they argue that features related to security (i.e. features 1-6) might increase the PCP strength and the features related to attracting users (i.e. features 7-9) might decrease PCP strength.

All their comparisons use the median as measure of central tendency for the strength values. They find that none of the security related features have an effect on PCP strength. However, all features related to attracting users have the anticipated effect, i.e. websites that either advertise themselves, display advertisements of third party websites, or those where users can choose alternatives are more likely to have weaker PCPs.

3. RELATED WORK

Aside from Florêncio and Herley's study, several different aspects of PCPs have been the focus of other studies. In the following, we present selected related research.

Komanduri et al. [9] investigated the security and usability properties of five different constructed (i.e. not taken from the wild) PCPs in an online study. They found that relatively simple password composition policies like mandating at least 16 characters and no other restrictions (i.e. their basic16 policy) yield much better security and usability results than mandating a length of at least 8 characters as well as the usage of uppercase letters, lowercase letters, numbers and symbols (i.e. their comprehensive8 policy).

Follow-up studies using similar constructed PCPs and focusing on the security [8] or both, security and usability, [18, 19] of PCPs could replicate these findings: the typical *comprehensive8* PCPs are among the weakest PCPs in terms of guessing resistance and also exhibit unfavourable usability properties. The authors recommend exchanging any *comprehensive8* PCP in use with one of three alternatives they identify in their study as exhibiting better security and better usability properties.

Focusing on PCPs used by websites in the wild, Kuhn and Garrison [10] conducted a survey of the difference in PCP

[1] `www.quantcast.com`

[2] Note the following two aspects of the original methodology, which were not explicitly mentioned in [5]: (1) if a password requires special characters, Florêncio and Herley [5] used a cardinality of 34 to calculate the PCP strength, regardless of the number of special characters actually allowed by the PCP, and (2) if a password allowed four character sets, but required only 3, they included letters instead of special characters in all instances.

strength over a period of two years. Their assessment of security was based solely on the minimum length of the password mandated by the PCP and therefore uses a much simpler measure than Florêncio and Herley [5] in their study. They found that more websites used PCPs in 2009 than in 2007 and that the mandatory length of passwords imposed by those PCPs increased in the same time frame.

Employing a similar methodology for the selection of their PCP sample as Florêncio and Herley, Seitz et al. [17] analysed the PCPs of the top 100 German websites (according to Alexa rankings). Their focus however, was to evaluate the potential of password reuse among these sites. They found that despite the great diversity among the PCPs, it is fairly easy to find a password that can be reused on virtually all websites.

Also considering the traffic rank of websites as an important factor, Preibusch and Bonneau [16] find that high traffic rank websites (determined by Alexa ranks) are more likely to attempt to prohibit password sharing among users by blocking listings of credentials at the password sharing community `bugmenot.com`.

Focusing on the practical aspects of password security, Florêncio et al. [6] summarise the password research relevant for system administrators. They present findings relating to the guessing resistance required to withstand offline and online attacks and also discuss implementation details such as appropriate hash functions. They advise to consider in the formulation of PCPs, that offline guessing attacks are much less frequent than originally thought and that online guessing attacks should be the focus when it comes to determining the guessing resistance of passwords.

In another work focused at practitioners, Zhang-Kennedy et al. [25] investigate long standing password management and composition rules. They discuss the viability of these rules based on the results of current research. Based on their findings, they introduce an updated set of password rules, aimed at decreasing the burden on users.

Aiming at increasing the security of existing PCPs in practice, Blocki et al. [2] describe a theoretical model for optimising password composition policies by maximising the PCP's minimum entropy from a set of sample passwords chosen by users.

4. METHODOLOGY

In conducting our replication of the original study by Florêncio and Herley [5], we computed the strength of the PCPs used in 2016 on those US websites used in the original study and re-investigated correlations between the websites' features as identified in [5]. We also applied their approach to a corresponding German sample. We chose a German sample in addition to the US ones for technical reasons: as Germans we can easily understand the PCP descriptions and conduct follow-up studies.

The purpose of our study is to investigate to which extent the results of the original study can be generalised across time and different countries, i.e. for which website features there exist differences between samples and where there exist none. Based on the results of our analyses, we answer the following questions:

RQ1: Has the average PCP strength in the US sample changed since the original study?

RQ2: Do the effects of the website features on the PCP strength from the original study still apply to the USA 2016 sample?

RQ3: How do the German and US samples compare in terms of PCP strength?

RQ4: Do the effects of the website features on the PCP strength from the original study translate to the German sample?

To ensure comparability with the original study, we employed the methodology as used by Florêncio and Herley [5] as closely as possible. However, in order to render our investigation viable and its results meaningful, we needed to adapt the methodology in some respects. In the following, we detail the alterations to the original study's methodology. Where not stated differently, we replicated the original methodology as described in section 2. All PCP data was collected in January 2016.

4.1 Identification of US Website Samples

For the US sample, we used the 75 websites from the original sample of Florêncio and Herley [5]. Five of these were excluded from our investigation due to the following reasons: *highschoolsports.net* was no longer available at the time of the survey, *youtube.com* now uses the google.com user account system, *ask.com* as well as *hollywood.com* seem no longer to have a user account system, and *typepad.com* did not provide information about its PCP on the website. Account creation would have required us to provide payment details. The remaining 70 websites were all included in our investigation. A list of all websites in the US sample can be found in the appendix in Table 5.

4.2 Identification of German Website Sample

We collected a comparative sample of 67 German websites. A list of all websites can be found in the appendix in Table 6. The websites were collected according to the categories defined in the original study: top, high, and medium traffic sites as well as banks, universities and government websites. We did not consider the miscellaneous category for the collection of the German sample due to its ambiguity. In the following, we describe the collection of the German website sample in detail.

German Top, High, and Medium Traffic Websites. Florêncio and Herley used the traffic rank information from the Quantcast service to identify top, high, and medium traffic websites. However, the service does not seem to provide reliable traffic rankings for German websites. While the site offers a list of the top 100 German websites, it does not seem to be representative of actual usage, e.g. none of the search engines included in popular browsers (Google, Bing, and Yahoo) appear. We thus used the alternative Alexa[3] rankings, since their list of German website rankings seemed much more representative and they have also been used in other pertinent studies (e.g. [17, 16]).

[3] `www.alexa.com/topsites/countries/DE`

Analogously to the original study, we used the ranks 1 to 20 and 101 to 110 for the top and high traffic categories respectively. Since the Alexa service only provides the 500 most visited websites, it was impossible to choose the ranks 1001 to 1010 for the medium traffic websites, as chosen in the original study. Therefore, we approximated by using the last ten ranks provided by Alexa[4]. As in the original study, websites without account systems were discarded.

German Banking Websites. For the banking category, we chose several of the largest banks in Germany [4, 11]. Unfortunately, some banks do not offer information about their policies on their websites. Therefore, we could only include four of the ten largest traditional German banks (i.e. banks with brick and mortar branch offices) and five of the ten largest German online-only banks (i.e. banks with no branch offices and only an online presence).

German University Websites. The German university websites used in this study represent the largest German universities, based on official government statistics [22]. The websites of the best-rated computer science departments are based on the CHE university ranking [24].

German Government Websites. Florêncio and Herley used the ten highest traffic websites with a `.gov` top-level domain. Germany does not have an equivalent to the US `.gov` domain, so we resorted to a different identification procedure. First, we manually identified the government websites on the Alexa list of the 500 most frequently-visited websites. This yielded only 5 government websites with an account system. To gather additional government websites for our sample we consulted a report comprising an extensive list of German government websites [14]. Using this list, we were able to identify an additional three websites with user account systems.

4.3 Identification of PCPs

The approach we took in order to identify the PCPs of both samples (USA 2016 and GER 2016) was similar to the one applied in the original study. Where possible, we created an account on the website[5]. Sometimes we used demo accounts to check the password policies. If neither of these approaches was possible, a web search was used to locate the PCP. For the US websites with multiple account systems, we used the PCPs for the same account systems as used in the original study, to support a meaningful comparison. On German websites with multiple account systems, we used the first PCP we found (analogously to the original study).

4.4 Website Accepts Advertising

We collected data regarding the placement of advertisements on the website. The original study relied on the advertising info provided by the Quantcast service to decide whether websites displayed advertisements from third party websites.

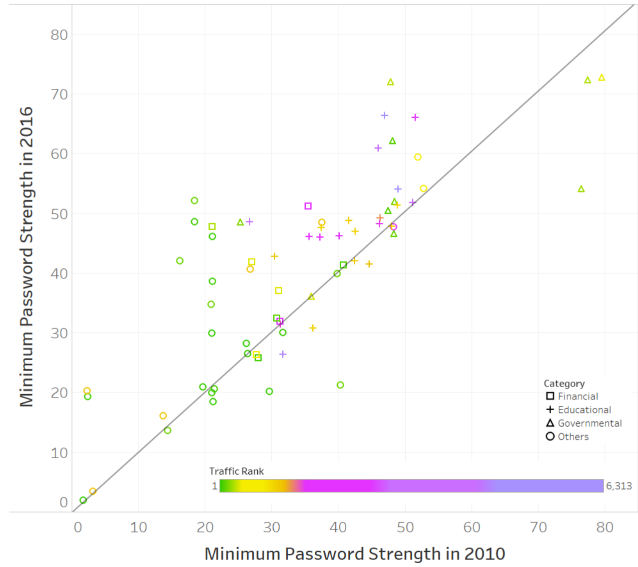


Figure 2: Scatterplot showing the difference in PCP strength (as determined by minimum password strength) for the US sample over time. Each point represents one website. The websites above the diagonal have adopted a stricter PCP since 2010, the websites below the diagonal have adopted a more lenient PCP.

This information is no longer available. We applied a manual approach. We visited each of the websites and navigated through the pages. If we found any advertisements, we categorised the respective website as having this feature, otherwise not.

5. RESULTS

In this section, we first present the results for the PCP's strength and then the results for the nine features[6] from the original study and their effects. Note that we refer repeatedly to *the three samples*, which relates to our German website sample (GER 2016) and our updated US website sample (USA 2016) as well as the US website sample from the original study (USA 2010). All correlations reported here were calculated using Pearson's correlation coefficient r using the R statistics system. Table 1 gives an overview of the investigated effects, the results of the original study, and the results of our own investigation.

5.1 Strengths of PCPs

The average strength of the US sample has grown significantly from 35.7 bits in 2010 to 41.4 bits in 2016, whereas the maximum values have declined from 79.0 bits to 71.5 bits respectively. The increase in average PCP strength is caused by 37 websites adopting a stronger PCP than in 2010, while only 8 websites adopted a weaker PCP. The remaining 25 websites did not change their PCP from 2010 to 2016. Figure 2 depicts the change in PCP strength over time in the US samples.

[4]Note, we can report Alexa ranks for the websites of the other categories in our sample, since it is possible to query the Alexa database for any website to get its rank. The identification of the medium traffic websites is hindered since the inverse is not possible (i.e. querying the Alexa database with a specific rank to get the corresponding website).

[5]Note, in some cases, it was not possible to create an actual account, but instead only to carry out the registration process. While no account was created, this process still gave us access to the PCP.

[6]We wont motivate the individual features in this work. Instead, we would like to refer the interested reader to the description in the original study [5] for a detailed description explaining the selection.

Table 1: Overview of the investigated website features and their hypothesised as well as actual effects on PCP strength. "↑" indicates an increase in strength, "↓" indicates a decrease in strength, "-" indicates no effect.

Website feature	Hypothesised effect on PCP strength [5]	Actual effect on PCP strength		
		USA 2010	USA 2016	GER 2016
Observation and evidence	↑	-	-	-
Size of the service		-	-	-
User name public		-	-	-
Value of the resources protected		-	-	-
Extractable value of the resources protected		-	-	-
Who lives with the consequences of a breach		-	-	-
Advertising accepted	↓	↓	↓	-
Site advertises		↓	-	-
User has choice		↓	↓	↓

Table 2: The median PCP strengths of the websites in the three samples. German websites generally employ weaker PCPs. In particular German banking websites stand out, exhibiting the lowest average in all three samples.

Sample	Overall	Traffic			Website type			
		Top	High	Medium	Bank	University	Government	Others
USA 2010	35.7	19.9	19.9	36.2	31.0	41.7	47.6	19.9
USA 2016	41.4	26.6	41.5	46.5	35.7	47.6	52.7	29.9
GER 2016	26.6	26.6	25.8	19.9	16.6	30.8	47.6	26.6

In comparison to the USA 2016 sample, the German 2016 sample shows a much smaller median of 26.6 bits and also a slightly lower maximum of 59.3 bits. The minimum is equal for all samples (3.3 bits), due to Wikipedia being present in all samples. Figure 1 depicts the distribution of the PCP strengths for all three samples. While the range decreased in the US samples from 2010 (75.7 bits) to 2016 (68.1 bits) and is even smaller for the German sample (55.9 bits), it remains very large in all samples.

Table 2 shows the differences between the samples for the different categories of websites distinguished in the original study. It becomes apparent, that the average PCP strengths have increased in the US sample for all categories. When comparing the the two samples from 2016, German websites employ on average weaker PCPs in every category. Figure 6 in the appendix illustrates the distributions of the PCP strength per category of all three samples.

5.2 Observation and Evidence

One might argue that website providers learn from past events and derive the PCPs from past experiences. Florêncio and Herley approached this question in the original study based on argumentation and while their arguments still hold today, we also applied an evidence-based approach in our investigation using information related to whether the US websites had been affected by a password-related breach or leak in the years between 2010 and 2016. To that end, we conducted web searches for each website in the US sample, in order to identify whether it had been affected by a password breach or leak since the original study. We used the Google search engine with the search terms "password breach", "password leak", "password hack", and "password incident", each in combination with the respective website's name. If we found a security incident exposing password data[7] on the first five pages of search results, we classified a website as having been victim of a breach or leak. Table 5 in the appendix shows the individual classification of each website in the US sample. This classification is, admittedly, only an approximation. Not all leaks are made public, which decreases the precision of our approach.

Using the classification, we split the websites into three categories: those having increased the the strength of their PCPs in the time between 2010 and 2016, those having reduced the strength of their PCPs in that time frame, and those with no change in their PCPs. We hypothesised that if websites operated on their past evidence, then websites that had been the target of a breach since 2010 would be more likely to have increased the strength of their PCPs. Table 3 shows the frequencies of websites classified as detailed above.

We conducted a Fisher's exact test to investigate the effect of websites being affected by a breach on the PCP strength of those websites. It yielded no significant results (FET: $p = 0.415$). Thus, the hypothesis that past breaches have an effect on PCP strength has to be rejected.

5.3 Size of the Service

Florêncio and Herley [5] hypothesised that PCP strength correlates with (a) the size of a website (as determined by the number of user accounts on that site) and (b) the traffic generated by the website (as determined by the Quantcast traffic rank, see section 2 for details). They reject these hypotheses based on the observation that top-traffic services

[7]Indirect attacks such as abusing reset mechanisms were not considered a breach or leak in the sense of our investigation since the actual password is not revealed when security questions are easily guessed. Also, attacks leaking passwords in the clear (such as phishing) were not considered, since stronger passwords do not protect against these kinds of attacks.

Table 3: Frequencies of websites classified along the two characteristics: (a) whether a website has been victim of a breach or not and (b) whether the website uses a stronger, weaker, or unchanged PCP.

	PCP is Stronger	Unchanged	Weaker
Breach	13	6	1
No breach	24	19	7

with many users (such as Facebook or GMail) have much weaker PCPs than universities which have significantly lower traffic ranks and also lower numbers of users (approximated from undergraduate enrolment). Since the traffic ranks in 2016 do not match the original sampling and in some cases deviate significantly from their 2010 ranks (e.g. Myspace had rank 16 in 2010 and ~1000 in 2016), a direct comparison is not possible.

However, since the original study, the top traffic websites have increased their number of users, e.g. Facebook from ~400 million to ~1700 million [21] or GMail from 91 million to 1000 million [20]. In contrast, in the same time frame the number of student enrolments remained steady for the lower traffic examples used by Florêncio and Herley [5], e.g. Ohio State University 51800 in 2010 and 51759 in 2016 [15]. Hence Florêncio and Herley's argument seems to hold. However, it must be acknowledged that the measure chosen by Florêncio and Herley for the approximation of the number of university user accounts (i.e. undergraduate student enrolments) might not be optimal (see section 7 for a discussion of this limitation).

In addition, we conducted a correlation analysis for our German sample based on the Alexa ranks. We found a weak negative correlation between the Alexa ranks of the websites in the German sample and the strength of their PCPs ($r = -0.16$). Thus, our results support the findings of the original study for the German sample as well.

5.4 User Name Public

When user names are publicly available, bulk guessing attacks, where attackers try only the most frequent passwords for all accounts known to them, become much more viable. Therefore, Florêncio and Herley hypothesised that websites with public user names might employ PCPs with a higher average strength. They assumed social networks', auction websites' and email providers' user names to be public and stated that for universities the user name is often public as well. For the US sample, the findings from the original study can be directly transferred: a Wilcoxon rank sum test results in rejecting the hypothesis that there is a difference in PCP strength between websites with public user names and websites where user names are not publicly accessible ($W = 648.5, p = 0.674$).

For the German sample it is of note that university user names do not seem to be public in general. Some universities from our sample use the student id number (e.g. FernUniversität Hagen), the student email address (e.g. Goethe Universität Frankfurt) or personal information such as first and last name (e.g. Universität Köln), and this might make bulk guessing easier. However, there are also universities in our sample which explicitly use random user names (e.g. TU Darmstadt or Universität des Saarlandes). In particular, for most German universities this information is not publicly available. However, even with this difference, a Wilcoxon rank sum test results in a rejection of the hypothesis that there is a difference in PCP strength between websites with public user names and websites where user names are not publicly accessible ($W = 456.5, p = 1.000$). Our findings fully support the findings of the original study for all samples.

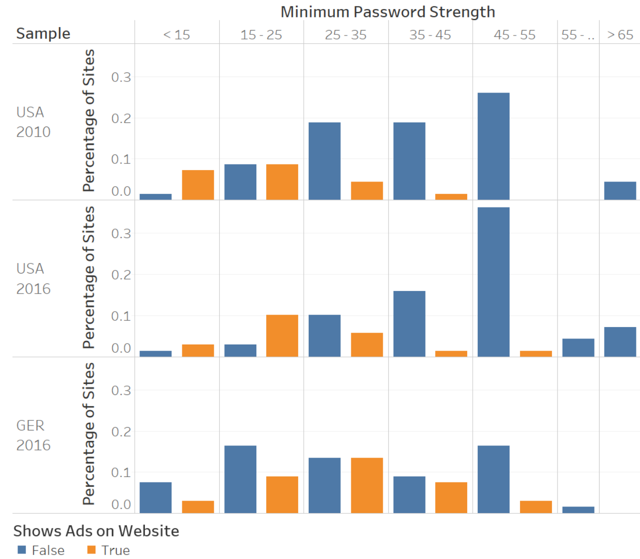


Figure 3: Histograms for the three samples, showing the distributions for websites that display third party advertisements and those that do not.

5.5 Value of the Resources Protected

While investigating the values of the resources protected in the USA 2016 sample, the same trend already described in the original study became apparent. Financial services have, on average, more lenient PCPs than government websites. In particular, the same example from the original study still holds. Both, Fidelity (increase in PCP strength since 2010) and Paypal (no difference in PCP strength since 2010), still have weaker PCPs than USAJobs (no difference since 2010). Table 2 shows the average PCP strengths of banking websites for all samples. German banking websites have the lowest average PCP strengths of all three samples (16.6 bits). US banking websites have significantly higher average PCP strengths in both 2010 and 2016.

5.6 Extractable Value of the Resources Protected

Florêncio and Herley hypothesise that the extractable value of user accounts might increase the PCP strength of respective websites. To identify the websites considered the most valuable, they consider those heavily targeted by phishers, since they argue that these represent the websites whose accounts offer the best monetisation. In contrast to their hypothesis, they find that the most phished brands in 2009 all have relatively low strength PCPs.

For the extractable value of the resources protected by the respective passwords, we see the same effects as described

Table 4: **The median PCP strengths of the websites in the three samples in relation to whether they accept third party advertisements, whether the websites advertise themselves, and whether the user can choose alternative websites.**

Sample	Accepts ads		Advertises		User choice	
	Yes	No	Yes	No	Yes	No
USA 2010	19.9	41.1	31.0	35.7	19.9	41.6
USA 2016	19.9	47.6	47.6	41.4	26.6	47.6
GER 2016	26.6	26.6	22.9	26.6	26.2	31.0

in the original study. According to the APWG [1], financial websites are still among the ones most heavily targeted by phishers. Close to 19% of phishing attacks target this sector. In contrast, less than 2% of attacks target government and education websites, both of which have much higher average PCP strengths.

5.7 Who Lives with the Consequences of a Breach

When a service has to compensate users for possible consequences, this financial threat could be a reason for website providers to enforce stronger PCPs. As noted in the original study, this was not the case in 2010. Our investigation provides even more evidence in this regard. Banks are still among the websites employing weak PCPs in all samples (in particular in the German sample). Yet they often compensate users for unauthorised transactions [5]. In Germany, account holders only have to cover the first 150€ themselves. Our investigation fully supports the original study's findings.

5.8 Advertising Accepted

Some websites generate their revenue through third party advertisements. Table 4 shows the median PCP strengths for the websites displaying third-party advertisements and the ones that do not. The US samples underline the findings of the original study: a Wilcoxon rank sum test indicated that websites displaying third-party advertisements had significantly weaker PCPs than those that did not display advertisements in the USA 2016 sample ($W = 759.0, p < 0.001$). It is interesting to note that the median PCP strengths for the websites accepting advertising did not change from 2010 to 2016. The overall increase in the USA 2016 sample stems solely from websites not displaying advertisements.

However, for the German sample, displaying adverts does not seem to have a significant effect on the average PCP strength. The median values for both groups of websites are the same as the overall median strength of 26.6 bits already reported in section 5.1. Consequently, a Wilcoxon rank sum test results in a rejection of the hypothesis that there is a difference in PCP strength between websites displaying third-party adverts and those that do not display third-party adverts in the German sample ($W = 554.5, p = 0.617$). Figure 3 illustrates this effect across all three samples.

5.9 Site Advertises

To generate traffic, some websites place advertisements on other websites. As an indicator of whether websites place such ads, we use (analogously to the original study) Google sponsored links. Figure 4 shows the distributions of PCP strength of websites utilising Google sponsored links and those that do not do this for all three samples. Again, there is no visible effect for this feature in the German sample; both median PCP strength values are almost identical at 26.6 bits for the non-advertising sites and 26.3 for the advertising sites (cf. Table 4). A Wilcoxon rank sum test supports this finding, resulting in a rejection of the hypothesis that there is a difference between advertising and non-advertising websites in the German sample ($W = 505.0, p = 0.366$).

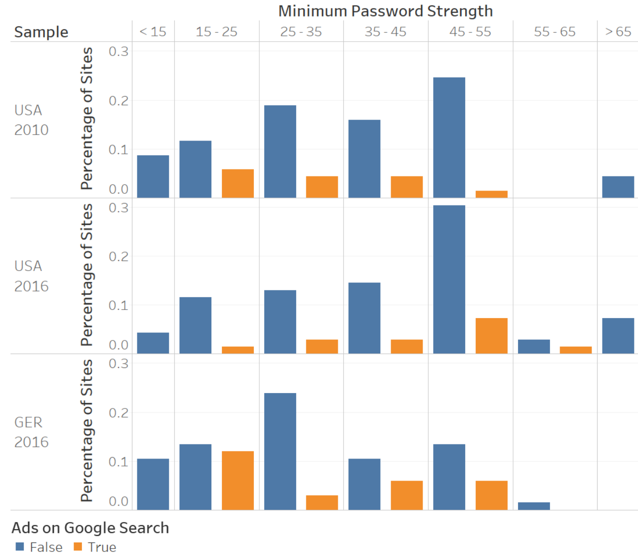


Figure 4: Histograms for the three samples, showing the distributions for websites that are advertising using sponsored links on Google Search and those that do not advertise using sponsored links.

In the USA 2016 sample, a Wilcoxon rank sum test results in rejecting the hypothesis that there is a difference in PCP strength between advertising and non-advertising websites as well ($W = 297.0, p = 0.667$). However, while not significant, the results might indicate a weak reversal effect, illustrated by Figure 4: in 2010, non-advertising websites had the higher average PCP strength, in 2016 the advertising websites have the higher average PCP strength (cf. Table 4).

5.10 User Has Choice

Concerning whether the user has a choice to use the website, the results are consistent for all three samples. Wilcoxon rank sum tests indicated that websites where users can choose between alternatives have significantly weaker PCPs than those where users have no choice for both, the USA 2016 sample ($W = 976.5, p < 0.001$) as well as the German sample ($W = 780.0, p = 0.004$). When comparing the two samples from 2016, German websites without alternatives for the user, unsurprisingly, have a lower PCP strength than the corresponding US websites. On the other hand, there is no difference between the websites of both samples where users can choose alternatives. Thus, the feature seems to have a similar effect on the strength of the PCPs in all three samples.

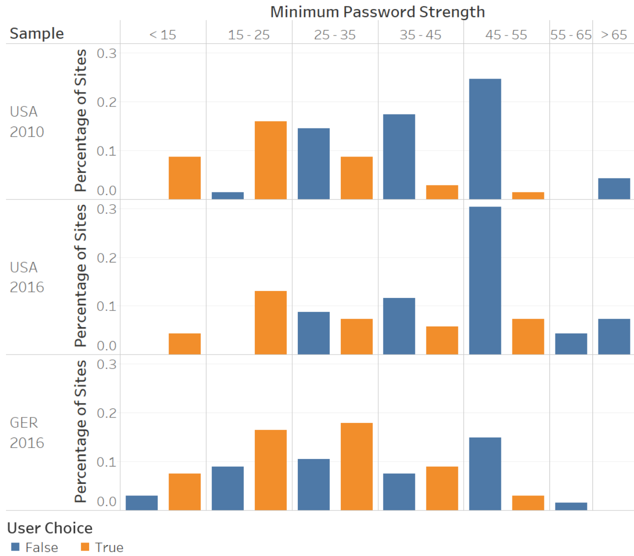


Figure 5: Histograms for the three samples, showing the distributions for websites where users can choose an alternative website and where users have no choice.

6. DISCUSSION

In our replication study, we re-investigated the effects of several website features on the strength of websites' PCPs. The goal of this replication was not only to revisit the original research questions of Florêncio and Herley [5], but to explore whether a comparison of PCPs over time and across country borders yields new findings. We discuss the findings related to each of our research questions in the following.

6.1 Has the average PCP strength in the US sample changed since the original study?

Based on the data from our replication of Florêncio and Herley's study [5], the answer to the first research question is a definitive "yes". The average PCP strength in the US sample has risen from 35.7 bits to 41.4 bits since the original study. 52.9% of the websites use 2016 stronger PCPs than they did in 2010. In contrast, only 11.4% of the websites used weaker PCPs.

While this trend supports similar findings by Kuhn and Garrison [10], it contradicts established expert opinion: NIST's newly drafted rules regarding password security [7] recommend using a PCP with *"at least 8 characters"* and *"no other complexity requirements for memorized secrets"*. Such a PCP has a minimum strength of 26.56 bits. The US average from 2016 is at 41.1 bits 35.8% higher than this recommendation. Therefore, it seems that PCPs found in the wild are much more complex (as determined by Florêncio and Herley's measure) than what is recommended. Such overly complex PCPs might be, usability-wise, alarming.

The reasons for this rise in PCP strength, however, cannot be identified from the data collected in our study. As in the original study, all hypotheses regarding factors increasing PCP strength had to be rejected. Thus, no explanation for the rise in PCP strength emerges from the original hypotheses. Furthermore, our additional investigation into the effects of breaches on PCP strength reveals that the observable rise in PCP strength between 2010 and 2016 cannot be attributed to the website being breached either. Consequentially, other features must be the driving force behind the rise in PCP strength. While it might be that website providers try to counter increasing attacker capabilities for offline attacks by employing stronger PCPs, this would contradict Florêncio et al's . [6] recommendations to focus on online guessing when designing PCPs, which is better addressed with lock-out policies. Yet the identification of the influential features, in this regard, constitutes one important focus for future work.

6.2 Do the effects of the website features on the PCP strength from the original study still apply to the USA 2016 sample?

Regarding the answer to our second research question, the results from our study concur with the findings of the original study for all features except one. All website features, which Florêncio and Herley hypothesised to increase PCP strength, still do not have that effect. Also, websites that display third party advertisements and websites where users can choose alternatives still have significantly weaker PCPs.

The only divergence from the findings in the original study is related to whether websites advertise to attract users. This feature seems to have lost its effect on the PCP strength. The reasons for this divergence, however, remain unclear.

6.3 How do the German and US samples compare in terms of PCP strength?

With respect to the third research question, the answer we can give from the results of our study is that websites in the German 2016 sample employ in every category on average weaker PCPs than those in the USA 2016 sample. For the three categories medium-traffic, banking, and education, the websites in the German 2016 sample have even lower average PCP strengths than the websites of those categories in the US sample had in 2010. Especially German banking websites stand out in this regard: While the passwords on these websites protect the most (monetary) value, they are created under the PCPs exhibiting the lowest average strength across all three samples.

However, one important aspect regarding German banking websites is that they implement two-factor transaction authorisation. The notable difference to two-factor authentication is that users can log in (authenticate) without the second factor, but authorising transactions requires a second factor (usually a so-called TAN, a transaction number for one-time use delivered either in advance e.g. as a list of TANs via mail or nowadays on demand e.g. via smartphone apps). Thus, the actual extraction of resources requires more than mere knowledge of the password, but carrying out this kind of attack is not impossible [12]. To gain further insight, we contacted a local bank. Their perspective is that tight lock-out policies and high security data centres made strong PCPs unnecessary. Hence, the trade-off of employing a lower-strength PCP in conjunction with tight lock-out policies and two-factor authorisation might be, usability-wise, a favourable trade-off. Whether users agree with this perspective and find such lower strength PCPs adequate for protection in the banking context remains an open question that we cannot answer without further investigation.

6.4 Do the effects of the website features on the PCP strength from the original study translate to the German sample?

The results of our study indicate that the effect of only one feature translates to the German sample. User choice is the only feature affecting the PCP strength in the German sample. When a user can choose alternatives to a certain website, that website is more likely to employ a weaker PCP. However, in contrast to the US samples, the display of adverts does not seem to have a significant effect on PCP strength in the German sample. Neither the display of advertisements, nor using adverts to attract users, has a significant effect on the PCP strength of websites in the German sample. Our data does not suggest any explanations for this. Regarding the effects of features hypothesised to increase PCP strength, all hypotheses had to be rejected. This confirms the findings of the original study.

Therefore, we argue that only two factors truly influence the strength of a website's PCP across the three samples: (1) a general tendency to enforce PCPs which are as strong as possible, and (2) the dependence on usability to attract users, leading to weaker policies. As already pointed out by Florêncio and Herley in the original study, and further supported by the findings of this replication, this trade-off is decided by websites more or less off-the-cuff. This holds for the US and German samples. To illustrate for the US sample: the range among US universities is 39.2 bits in 2016. However, it is unclear, why Princeton (PCP strength of 65.8 bits) should feel the need to enforce a significantly stronger PCP than Northwestern University (PCP strength of 26.6 bits). With respect to the German sample, the average strength of the PCPs seems, with 26.6 bits, to be very close to NIST's recommendation (26.56 bits). However, the large range of 55.9 bits across PCP strengths and, in particular, high strength PCPs on seemingly low value sites (e.g. 47.6 bits for the news site `spiegel.de`) give rise to doubt regarding a more systematic approach being applied to PCP choice on German websites.

7. LIMITATIONS

As already acknowledged by Florêncio and Herley [5], the minimum PCP strength measure employed in their study can only serve as a rough estimation and more precise measures of guessing resistance exist. However, as Florêncio and Herley pointed out in the original study, their measure is not intended to model resistance to guessing attacks, but only complexity of the resulting passwords. Since we adopt this measure to perform our replication, this limitation applies to our study as well.

We also decided to not use any additional measures of password security, since any reliable estimate of a PCPs strength (e.g. α-guesswork [3] or guess numbers [23]) would require collecting passwords created under the respective PCP. However, collecting adequate numbers of passwords for the calculation of these measures is beyond the scope of this work.

Another limitation that arises from adopting the original methodology and the nature of performing a replication is that PCP strength is only investigated in relation to the website features. The effects of other influencing factors such as technologies employed by the user (e.g. two-factor authentication, password managers, etc.) are not considered. Especially, two-factor authentication might play a role in some categories: as explained before in section 6.3 German banks require the user to provide a second factor to authorise transactions. It must be assumed that this influences the PCP choice of banks. However, other factors might play a role here as well. For example, traditional banks (i.e. banks with brick and mortar branch offices) might have relatively strict lockout policies, since their customers can simply visit the local branch office to get their account unlocked. Therefore, while future study designs should include these interesting extensions of the methodology and consider such technologies, it was beyond the scope of this replication study.

The third limitation of our replication study is also shared with the original study. For the identification of the PCPs in our samples, we followed the same methodology as Florêncio and Herley. Thus, we also created an account at the website whenever possible (the information whether an account was created is available for each of the websites in our samples in tables 5 and 6 in the appendix). However, when this was not possible we also followed the methodology of the original study and conducted a web search. This can lead to imprecisions in the samples, since sometimes the found PCPs might represent guidelines not enforced during the actual password choice or for a university might only be enforced for a specific account system, but not for others.

The final limitation that our replication study shares with the original study is the approximation of user accounts at universities by undergraduate enrolment numbers. Using purely these numbers might not be optimal, since the number of accounts managed by universities nowadays might only loosely correlate with the number enrolled undergraduate students due to the emergence of other account systems at the universities (e.g. affiliated research institutes, alumni, donors, or even accounts for the purchase of sports tickets)[8]. However, we argue that this does not affect our results, since even if the undergraduate enrolment is not fully representative of the number of user accounts at universities, it is unlikely that universities reach the numbers of users of the top traffic websites.

A limitation arising from the longitudinal analysis is that we decided to use the same websites as the original study for the USA 2016 sample instead of collecting a new sample from the same categories. Therefore, some of the websites now belong to a different category. However, we decided to use the same website, since this affects less than 1/10 of the sample and we believe the longitudinal comparison (enabled only by using the same websites) adds special value to this paper.

Lastly, the additional breach analysis we conducted (cf. section 5.2) should only be treated as an approximation. While some countries have passed laws mandating the reporting of data breaches (cf. e.g. [13]), this does not hold for all jurisdictions and consequentially not all leaks are made public. Moreover, our search terms might have been insufficient to identify all available information on breaches at the respective sites.

8. CONCLUSION

In this paper, we presented a replication of the study by Florêncio and Herley [5]. Thereby, the contribution of our

[8]Thanks to reviewer 1 for pointing this out.

paper is twofold: (1) the comparison of password composition policies in the US sample over time, and (2) the comparison of password composition policies across country borders (i.e. between Germany and the US).

Regarding the first contribution, it became apparent that US PCPs have become, on average, stronger and that all but one website feature have retained their effects on PCP strength in the intervening years. While the former is in line with the findings of similar studies [10], it contradicts established expert opinion and might be, in terms of usability, an alarming finding. Moreover, our results indicate that two website features correlate with decreased PCP strength in the USA 2016 sample (i.e. "advertising accepted" and "user has choice"), but none of the website features seem to correlate with increased PCP strength in practice. Therefore, future work is needed to identify the reasons behind the rise in PCP strength in the US from 2010 to 2016. With respect to the effects of the website features on PCP strength in the US samples, only the effect associated to whether websites advertise to attract users seems to have changed. The effect could not be found in the 2016 sample.

Regarding the second contribution, we observed, on average, lower PCP strengths in the German sample than in the US samples. German banks stand out as having particularly weak PCPs. Together with the fact that "User has choice" emerges as the only website feature exhibiting an effect in the German sample, it seems that German banks are especially keen to maximise usability and optimise the user experience. They provide the user with a favourable trade-off by combining tight lock-out policies with the requirement of a second factor to authorise transactions. However, whether users consider this trade-off adequate for the banking context, or might even want to make similar trade-offs in other contexts as well, is open for future investigation.

9. ACKNOWLEDGMENT

We would like to thank Karen Renaud, Benjamin Reinheimer and Annika Hilt for their valuable input and feedback. Furthermore, we thank the anonymous reviewers and our shepherd Mary Ellen Zurko for their construcutve and helpful feedback.

This work has been developed within the project 'KMU AWARE' which is funded by the German Federal Ministry for Economic Affairs and Energy under grant no. BMWi-VIA5-090168623-01-1/2015. Furthermore, it has been supported in part by the German Federal Ministry of Education and Research (BMBF) within CRISP (www.crisp-da.de/). The authors assume responsibility for the content.

10. REFERENCES

[1] Anti-Phishing Working Group. Phishing Activity Trends Report - 1st Quarter 2016. Technical report, 2016.

[2] J. Blocki, S. Komanduri, A. Procaccia, and O. Sheffet. Optimizing password composition policies. In *EC '13: Proceedings of the fourteenth ACM conference on Electronic commerce*. ACM Request Permissions, June 2013.

[3] J. Bonneau. The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords. In *IEEE Symposium on Security and Privacy*, pages 538–552, 2012.

[4] Bundesverband deutscher Banken e.V. Zahlen, Daten, Fakten der Kreditwirtschaft. `https://bankenverband.de/publikationen/zahlen-daten-fakten/`. Accessed: 2016-01-16.

[5] D. Florêncio and C. Herley. Where do security policies come from? In *SOUPS '10: Proceedings of the Sixth Symposium on Usable Privacy and Security*, page 1. ACM Press, 2010.

[6] D. Florêncio, C. Herley, and P. C. van Oorschot. An Administrator's Guide to Internet Password Research. In *Large Installation System Administration Conference*, pages 35–52, 2014.

[7] P. A. Grassi, J. L. Fenton, E. M. Newton, R. A. Perlner, A. R. Regenscheid, W. E. Burr, J. P. Richer, N. B. Lefkovitz, J. M. Danker, Y.-Y. Choong, K. K. Greene, and M. F. Theofanos. NIST Draft SP 800-63B: Digital Identity Guidelines - Authentication and Lifecycle Management. `https://pound.netzpolitik.org/wp-upload/Analyse-staatlicher-Websites-Bewertung.pdf`. Accessed: 2016-01-05.

[8] P. G. Kelley, S. Komanduri, M. L. Mazurek, R. Shay, T. Vidas, L. Bauer, C. Wiedeman, L. F. Cranor, and J. 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.

[9] S. Komanduri, R. Shay, P. G. Kelley, M. L. Mazurek, L. Bauer, C. Wiedeman, L. F. Cranor, and S. Egelman. Of Passwords and People: Measuring the Effect of Password-Composition Policies. In *CHI '11: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2595–2604, New York, New York, USA, 2011. ACM Press.

[10] B. T. Kuhn and C. Garrison. A survey of passwords from 2007 to 2009. In *Information Security Curriculum Development Conference*, pages 91–94, New York, New York, USA, Sept. 2009. ACM.

[11] Modern Banking. Die größten Direktbanken gemessen an der Kundenzah. `http://www.modern-banking.de/marktanteil_direktbanken.htm`. Accessed: 2016-01-16.

[12] C. Mulliner, R. Borgaonkar, P. Stewin, and J.-P. Seifert. SMS-Based One-Time Passwords: Attacks and Defense. In *Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 150–159. Springer, 2013.

[13] National Conference of State Legislatures. Security Breach Notification Laws. `http://www.ncsl.org/research/telecommunications-and-information-technology/security-breach-notification-laws.aspx`. Accessed: 2017-02-22.

[14] Netzpolitik.org. Trackingtools auf Websites staatlicher Institutionen. `https://pound.netzpolitik.org/wp-upload/Analyse-staatlicher-Websites-Bewertung.pdf`. Accessed: 2016-01-20.

[15] Ohio State University Institutional Research and Planning. Statistical Summary. `https://www.osu.edu/osutoday/StatisticalSummary2015.pdf`, 2015.

[16] S. Preibusch and J. Bonneau. The Password Game:

Negative Externalities from Weak Password Practices. In *International Conference on Decision and Game Theory for Security*, pages 192–207, Berlin, Heidelberg, Nov. 2010. Springer, Berlin, Heidelberg.

[17] T. Seitz, M. Hartmann, J. Pfab, and S. Souque. Do Differences in Password Policies Prevent Password Reuse? In *CHI Conference Extended Abstracts*, pages 2056–2063, New York, New York, USA, May 2017. ACM.

[18] R. Shay, L. F. Cranor, S. Komanduri, A. L. Durity, P. S. Huh, M. L. Mazurek, S. M. Segreti, B. Ur, L. Bauer, and N. Christin. Can long passwords be secure and usable? *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*, 2014.

[19] R. Shay, S. Komanduri, A. L. Durity, P. S. Huh, M. L. Mazurek, S. M. Segreti, B. Ur, L. Bauer, N. Christin, and L. F. Cranor. Designing Password Policies for Strength and Usability. *ACM Transactions on Information and System Security (TISSEC)*, 18(4):13–34, 2016.

[20] statista. Number of active Gmail users worldwide from January 2012 to February 2016 (in millions). `https://www.statista.com/statistics/432390/active-gmail-users/`. Accessed: 2016-02-29.

[21] statista. Number of monthly active Facebook users worldwide (in millions). `https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/`. Accessed: 2016-02-29.

[22] Statistisches Bundesamt. Hochschulen. `https://www.destatis.de/DE/ZahlenFakten/GesellschaftStaat/BildungForschungKultur/Hochschulen/Hochschulen.html`. Accessed: 2016-01-17.

[23] B. Ur, S. M. Segreti, L. Bauer, N. Christin, L. F. Cranor, S. Komanduri, D. Kurilova, M. L. Mazurek, W. Melicher, and R. Shay. Measuring real-world accuracies and biases in modeling password guessability. In *USENIX Security Symposium*, 2015.

[24] Zeit Campus. CHE Hochschulranking 2015/16. `http://ranking.zeit.de/che2015/en/`. Accessed: 2016-01-17, now replaced with the 2016/17 version: `http://ranking.zeit.de/che2016/en/`.

[25] L. Zhang-Kennedy, S. Chiasson, and P. van Oorschot. Revisiting Password Rules: Facilitating Human Management of Passwords. In *Information Assurance and Security Workshop*, 2016.

Table 5: The US website sample (USA 2016) comprising 70 websites. Traffic ranks according to Quantcast.

Website	Traffic Rank	Account Created?.	Min. Length	Size Charset	Min. Strength 2016	Min. Strength 2010	Accepts Ads?	Places Ads?	User Choice	Affected by Breach
Top Traffic Sites										
Google	1	y	8	10	26,6	26,6	y	n	y	n
Facebook	2	y	6	10	19,9	19,9	y	n	y	n
Yahoo	3	y	9	10	29,9	19,9	y	n	y	y
AOL	6	y	8	10	26,6	26,6	y	n	y	y
Live	8	y	8	36	41,4	19,9	y	n	y	y
Wikipedia	9	y	1	10	3,3	3,3	n	n	y	n
eBay	10	y	6	36	31,0	31,0	y	n	y	y
Amazon	11	y	6	10	19,9	19,9	y	y	y	y
weather	13	y	6	10	19,9	19,9	y	n	y	n
answers	15	y	6	10	19,9	3,3	y	n	y	n
Myspace	16	y	6	10	19,9	31,0	n	n	y	n
Craigslist	17	y	8	26	37,6	19,9	n	n	y	n
adobe	20	y	8	62	47,6	19,9	n	y	y	y
High Traffic Sites										
nih.gov	101	n	8	62	47,6	47,6	n	n	n	n
capitalone.com	102	n	8	36	41,4	41,4	n	y	n	n
rockyou.com	103	y	6	10	19,9	41,4	y	n	y	n
overstock.com	107	y	8	36	41,4	16,6	n	n	y	n
latimes.com	108	y	7	36	36,2	19,9	y	n	y	n
intuit.com	109	y	8	96	52,7	19,9	n	y	y	y
cbssports.com	110	y	4	10	13,3	13,3	y	n	y	n
Medium Traffic Sites										
wowwiki.com	1001	y	1	10	3,3	3,3	y	n	y	n
virginia.edu	1002	n	8	62	47,6	36,2	n	n	n	n
pgatour.com	1003	y	6	10	19,9	3,3	y	n	y	n
mit.edu	1006	n	8	36	41,4	31,0	n	n	n	n
okcupid.com	1007	y	5	10	16,6	13,3	y	n	y	n
istockphoto.com	1008	y	8	36	41,4	25,8	n	y	y	n
Banks										
Fidelity	224	n	8	62	47,6	19,9	n	y	n	n
Vanguard	629	n	8	10	26,6	26,6	n	n	n	n
Schwab	2266	n	6	36	31,0	31,0	n	y	n	n
WellsFargo	80	n	6	36	31,0	31,0	n	n	n	n
BoA	48	n	8	36	41,4	41,4	n	n	n	n
J P Morgan Chase	2186	n	8	86	51,4	36,2	n	n	n	n
Citibank	316	n	6	62	35,7	31,0	n	n	n	n
PayPal	29	y	8	10	26,6	26,6	n	n	y	n
US Bank	316	n	8	36	41,4	26,6	n	n	n	n
Large Universities										
Ohio State U	1811	n	8	62	47,6	41,4	n	y	n	n
Arizona State U	3288	n	10	62	59,5	47,6	n	y	n	y
U. of Florida	1382	n	8	62	47,6	47,6	n	n	n	n
U. of Minn.	919	n	6	36	31,0	35,7	n	n	n	n
U. of Texas	946	n	8	62	47,6	47,6	n	n	n	n
U. of Central Florida	6313	n	8	96	52,7	47,6	n	n	n	y
Michigan State U.	1174	n	8	62	47,6	47,6	n	n	n	n
Texas A & M	1418	n	8	62	47,6	35,7	n	n	n	y
U. South Florida	2364	n	8	62	47,6	35,7	n	n	n	n
Penn. State U.	977	n	8	36	41,4	41,4	n	n	n	y
Universities with top CS departments										
MIT	1006	n	8	36	41,4	31,0	n	n	n	n
Stanford	858	n	8	96	52,7	47,6	n	n	n	y
UC Berkeley	905	n	9	36	46,5	41,4	n	n	n	n
CMU	3651	n	8	96	52,7	52,0	n	n	n	n

Continued on next page

Table 5 – continued from previous page

Website	Traffic Rank	Account Created?.	Min. Length	Size Charset	Min. Strength 2016	Min. Strength 2010	Accepts Ads?	Places Ads?	User Choice	Affected by Breach
UIUC	3384	n	8	62	47,6	26,1	n	n	n	n
Cornell	955	n	8	62	47,6	41,7	n	n	n	y
Princeton	1879	n	10	96	65,8	52,7	n	n	n	y
U. of Washington	1032	n	8	36	41,4	45,6	n	n	n	n
Georgia Tech.	4687	n	11	62	65,5	47,6	n	n	n	n
U. of Texas	946	n	8	62	47,6	47,6	n	n	n	y
Government Sites										
irs.gov	63	n	8	70	49,0	47,6	n	n	n	y
usps.com	68	y	10	75	62,3	47,6	n	n	n	n
nih.gov	101	n	8	62	47,6	47,6	n	n	n	n
ca.gov	124	n	8	96	52,7	47,6	n	n	n	n
ed.gov	141	y	8	62	47,6	26,6	n	n	n	n
noaa.gov	199	n	8	96	52,7	77,1	n	n	n	n
weather.gov	228	n	12	62	71,5	77,1	n	n	n	n
census.gov	246	n	12	62	71,5	47,6	n	n	n	y
ssa.gov	276	n	7	36	36,2	36,2	n	n	n	n
nasa.gov	342	n	12	62	71,5	79,0	n	n	n	y
Other Sites										
U. of Phoenix	873	y	8	62	47,6	36,2	n	y	y	n
Columbia	1350	n	6	36	31,0	31,0	n	y	n	n
Northwestern	4457	n	8	10	26,6	31,0	n	n	n	n
VA	558	n	9	96	59,3	52,7	n	n	n	n
USAJobs	590	y	8	96	52,7	52,7	n	n	y	n
TreasuryDirect	2421	n	8	70	49,0	47,6	n	n	n	y
Twitter	31	y	6	10	19,9	19,9	n	n	y	y

Table 6: The German website sample (GER 2016) comprising 67 websites. Traffic ranks according to Alexa.

Website	Traffic Rank	Account created?	Min. Length	Size Charset	Min. Strength	Accepts Ads?	Places Ads?	User Choice
Top Traffic Sites								
Google.de	1	y	8	10	26,6	y	n	y
Amazon.de	2	y	6	10	19,9	y	y	y
Facebook.com	3	y	6	10	19,9	y	n	y
Ebay.de	5	y	6	36	31,0	y	n	y
Wikipedia	7	y	1	10	3,3	n	n	y
Web.de	8	y	8	10	26,6	y	n	y
Ebay-kleinanzeigen.de	9	y	6	10	19,9	y	n	y
T-online.de	10	y	8	36	41,4	y	n	y
Gmx.net	11	y	8	10	26,6	y	n	y
Bild.de	13	y	6	10	19,9	y	n	y
Yahoo.com	14	y	9	10	29,9	y	n	y
Spiegel.de	15	y	8	62	47,6	y	n	y
Xhamster.com	17	y	4	10	13,3	y	n	y
Paypal.com	18	y	8	10	26,6	n	n	y
Focus.de	19	y	8	10	26,6	y	n	y
Live.com	20	y	8	36	41,4	y	n	y
High Traffic Sites								
Mytoys.de	101	y	5	36	25,8	y	y	y
vodafone.de	102	y	8	36	41,4	n	y	n
aol.com	103	y	8	10	26,6	y	n	y
zdf.de	104	y	1	10	3,3	n	n	y
netflix.com	105	y	4	10	13,3	n	n	y
duden.de	106	y	6	96	39,5	y	n	y
eventim.de	107	y	5	10	16,6	n	y	y
xvideos.com	109	y	8	26	37,6	y	n	y
bonprix.de	110	y	6	10	19,9	n	y	y
Medium Traffic Sites								
proxer.me	491	y	8	10	26,6	y	n	y
Auto-motor-und-sport.de	493	y	5	10	16,6	y	n	y
pcgames.de	494	y	8	36	41,4	y	n	y
etsy.com	495	y	6	10	19,9	n	y	y
netdoktor.de	496	y	1	10	3,3	y	n	y
opodo.de	497	y	7	36	36,2	y	y	y
clipfish.de	499	y	5	10	16,6	y	n	y
Banks								
Deutsche Bank	74	y	5	10	16.6	n	y	n
KfW	2017	y	8	62	47,6	n	y	n
NordLB	13002	y	5	10	16,6	n	n	n
Deutsche Postbank	37	y	5	10	16,6	n	y	n
Ing-diba	89	n	5	10	16,6	n	y	y
DKB	125	n	5	36	25,8	n	y	y
comdirect	132	y	7	36	36,2	n	y	y
Volkswagenbank	1816	y	8	10	26,6	n	y	y
Consorsbank	410	y	5	10	16,6	n	y	y
Large Universities								
FU Hagen	1728	n	8	62	47,6	n	y	n
LMU München	869	n	4	36	20,7	n	n	n
U Köln	979	n	6	36	31,0	n	n	n
Goethe U Frankfurt	1235	n	6	10	19,9	n	n	n
Ruhr U Bochum	4989	n	8	10	26,6	n	n	n
WWU Münster	957	n	8	62	47,6	n	y	n
RWTH Aachen	1031	n	6	10	19,9	n	n	n
U Hamburg	1278	n	8	62	47,6	n	n	n

Continued on next page

Table 6 – continued from previous page

Website	Traffic Rank	Account created?	Min. Length	Size Charset	Min. Strength	Accepts Ads?	Places Ads?	User Choice
U Duisburg-Essen	1139	n	8	62	47,6	n	n	n
FAU Erlangen-Nürnberg	2664	n	4	10	13,3	n	n	n
Universities with top CS departments								
RWTH Aachen	1031	n	6	10	19,9	n	n	n
U Augsburg	3355	n	8	36	41,4	n	n	n
Jacobs U Bremen	15549	n	6	10	19,9	n	y	n
U Magdeburg	3750	n	6	34	30,5	n	n	n
Hasso-Plattner-Inst. Potsdam	16678	y	1	10	3,3	n	n	n
U Bayreuth	2363	n	8	10	26,6	n	n	n
TU Darmstadt	1990	y	9	96	59,3	n	n	n
FAU Erlangen-Nürnberg	2210	n	4	10	13,3	n	n	n
U Konstanz	3656	n	8	36	41,4	n	n	n
U des Saarlandes Saarbrücken	2368	n	6	36	31,0	n	n	n
Government Sites								
bundestag.de	2101	n	8	62	47,6	n	n	n
arbeitsagentur.de	97	n	8	62	47,6	n	n	n
bundesregierung.de	3440	n	8	62	47,6	n	n	n
bund.de	436	n	8	62	47,6	n	n	n
destatis.de	2240	n	8	96	52,7	n	n	n
bayern.de	245	n	8	44	43,7	n	n	n
nrw.de	309	n	8	10	26,6	n	n	n
europa.eu	377	n	8	10	26,6	n	n	n

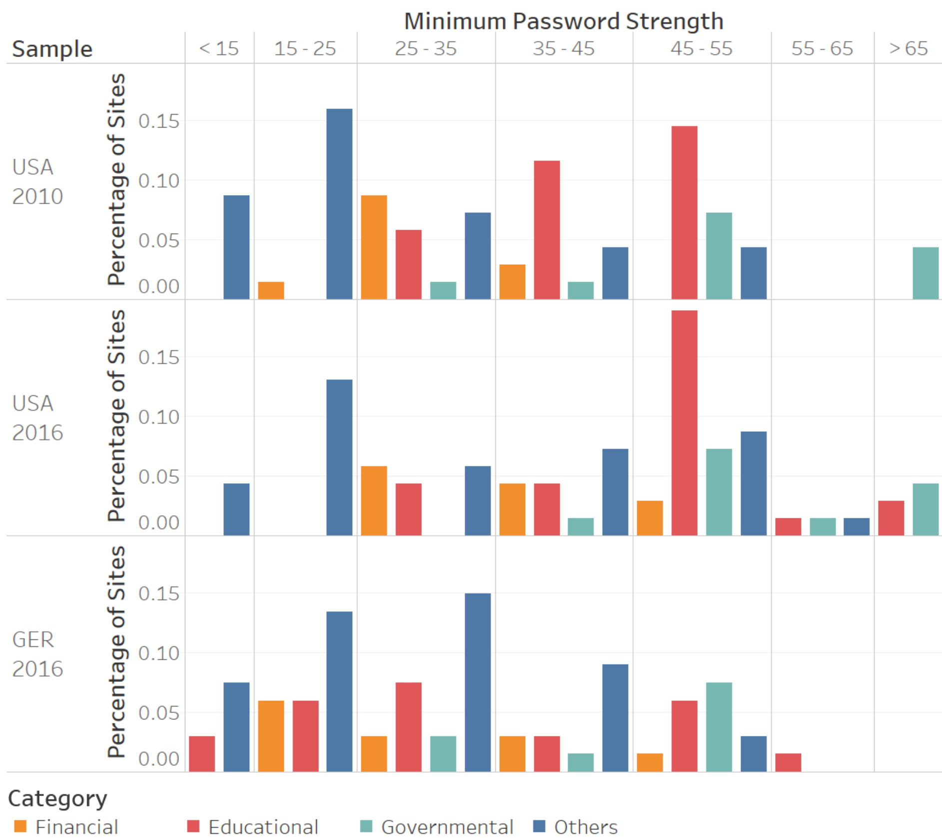


Figure 6: Histograms of the average PCP strength for all three samples along the different categories of websites.

Is that you, Alice? A Usability Study of the Authentication Ceremony of Secure Messaging Applications

Elham Vaziripour, Justin Wu, Mark O'Neill, Ray Clinton, Jordan Whitehead,
Scott Heidbrink, Kent Seamons, Daniel Zappala
Brigham Young University
{*elhamvaziripour,justinwu,mto,rclinton,jaw,sheidbri*}*@byu.edu*, {*seamons,zappala*}*@cs.byu.edu*

ABSTRACT

The effective security provided by secure messaging applications depends heavily on users completing an authentication ceremony—a sequence of manual operations enabling users to verify they are indeed communicating with one another. Unfortunately, evidence to date suggests users are unable to do this. Accordingly, we study in detail how well users can locate and complete the authentication ceremony when they are aware of the need for authentication. We execute a two-phase study involving 36 pairs of participants, using three popular messaging applications with support for secure messaging functionality: WhatsApp, Viber, and Facebook Messenger. The first phase included instruction about potential threats, while the second phase also included instructions about the importance of the authentication ceremony. We find that, across the three apps, the average success rates of finding and completing the authentication ceremony increases from 14% to 79% from the first to second phase, with second-phase success rates as high as 96% for Viber. However, the time required to find and complete the ceremony is undesirably long from a usability standpoint, and our data is inconclusive on whether users make the connection between this ceremony and the security guarantees it brings. We discuss in detail the success rates, task timings, and user feedback for each application, as well as common mistakes and user grievances. We conclude by exploring user threat models, finding significant gaps in user awareness and understanding.

1. INTRODUCTION

Recent disclosures of government surveillance and fears over cybersecurity attacks have increased public interest in secure and private communication. As a result, numerous secure messaging applications have been developed, including Signal, WhatsApp, and Viber, which provide end-to-end encryption of personal messages [19].

Most popular secure messaging applications are usable because they hide many of the details of how encryption is provided. Indeed, people are primarily using these applications due to peer influence, not due to concern over privacy or security [5].

The strength of the security properties of these applications rests on the *authentication ceremony*, in which users validate the encryption keys being used. Unfortunately, there is evidence that most users do not know how to successfully complete this ceremony and are thus vulnerable to potential attacks [15]. Any user who does not execute the authentication ceremony for a particular conversation is essentially trusting the application's servers to correctly distribute the encryption keys. This leaves users vulnerable to compromise threats that can intercept communications.

Several recent papers have shown that the authentication ceremony in secure messaging applications is difficult to use and prone to failure. A study of Signal showed that users, all of whom were computer science students, were highly vulnerable to active attacks [15]. A comparison of WhatsApp, Viber, Telegram, and Signal, found that most users were unable to properly authenticate [8], though after being instructed on what to do most users were subsequently able to authenticate after a key reset.

This state of affairs motivates our study, which examines to what extent users can successfully locate and complete the authentication ceremony in secure messaging applications if they are aware of the need for authentication. To answer this question, we conduct a two-phase user study of WhatsApp [22], Facebook Messenger [7], and Viber [21]. We chose these applications because of their popularity and their different designs. The authentication ceremony in WhatsApp uses either a QR code or a numeric key representation that users can compare. Viber presents a numeric key representation and provides functionality for users to call each other within the ceremony to compare the key. Facebook Messenger provides a numeric representation of the keys for both users. In addition to these differences, WhatsApp and Viber offer only secure messaging, while Facebook Messenger offers both insecure and secure messaging. We are curious as to whether the inclusion of an insecure messaging interface hinders the ability of users to find and successfully use secure messaging and the authentication ceremony.

In the first phase of our study, we asked 12 pairs of participants to complete a scenario where one participant needed to send a credit card number to the other participant. They were both instructed to verify that they were truly communicating with their partner (authenticity) as well as to ensure that no other party could read their messages (confidential-

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

ity). Participants were told the application would help them accomplish these goals.

In the second phase of the study, we presented 24 pairs of participants with the same task and scenario provided in the first phase. However, unlike the first phase, participants first read through an additional set of instructional slides before beginning the task. These slides informed them about traffic interception, that secure messaging applications use a “key” to secure conversations, and that to be secure they needed to confirm that they saw the same “key” as their partner. Participants were not instructed on how to use the applications to compare keys, nor shown any screenshots of the authentication ceremony; they were only told that each application had some way of providing this functionality. For both study phases, the method used for authentication was left to their discretion.

Each phase was a within-subjects study, and all participants engaged with all three applications in each phase. Participants differed between the two phases, allowing us to capture between-subjects differences in instruction between the two phases. We measured success rates in completing the authentication ceremony, time to both find and complete the ceremony, and user feedback on the applications, which includes System Usability Scale (SUS) scores, ratings of favorite application, ratings of trustworthiness for each application, and qualitative feedback.

Our findings include:

- In the first phase, despite the instruction about potential threats, the overall success rate over all participants and all applications was 14%, and only two of the twelve pairs of participants successfully located and completed the authentication ceremony. All other pairs attempted to authenticate one another through video calls, asking questions that required special knowledge to answer, or other ad hoc methods.

- In the second phase, the overall success rate increased to 79% for location and completion of the authentication ceremony. The success rates for the three applications were: 96% for Viber, 79% for WhatsApp, and 63% for Facebook Messenger.

- Viber’s higher success rate was statistically significant when compared to the other two applications. This is interesting because Viber’s authentication ceremony uses an in-app phone call and provides a UI that helps users view and read the encryption key during the phone call. Both WhatsApp and Facebook Messenger also provide manual verification of the encryption key, but do not provide this assistance. For both of these applications, numerous participants sent the key through in-app text, voice, and video, with a minority comparing the keys in person. Nearly half of participants chose to use the option WhatsApp provided for scanning a QR code.

- Averaged across the three applications, discovery of ceremony functionality took 3.2 minutes with ceremony completion necessitating another 7.8 minutes.

- All applications were rated in the “C” range on the System Usability Scale, indicating a need for significant usability enhancements.

- Most participants had not heard of Viber prior to their participation in our study. Trust ratings were very low in the first phase, but increased significantly in the second phase, when some instruction about security was received. This provides some evidence that learning about security features can enhance trust in a secure messaging application.

- Numerous participants complained about the length of the encryption key when having to compare it manually, taking shortcuts and often feeling fatigued by the process.

- Our qualitative data indicates that our participants have a healthy wariness for, and high-level understanding of: impersonation attacks, government and developer backdoors, and physical theft. They are, however, generally unaware of the existence of man-in-the-middle attacks, both passive and active. Our data is inconclusive on whether users make the connection between this ceremony and its security guarantees.

Our main takeaway is that even with an awareness of potential threats, users are not aware of and do not easily find the authentication ceremony in the secure messaging applications we tested. If given some instruction on the importance of comparing keys, they can find and use the authentication ceremony, and Viber’s second-phase success rate indicates that a high success rate is a realizable goal. However, for all applications, the time to find and use the authentication ceremony is unsatisfactory from a usability standpoint. The natural tendency of our participants to use personal characteristics for authentication, such as a person’s voice, face, or shared knowledge, indicates that future work could leverage this for a more user-understandable method of authentication.

2. RELATED WORK

Several papers have studied the usability of the authentication ceremony in secure messaging applications.

Two papers study the usability of the ceremony in a particular application. Schröder et al. studied Signal, showing that users were vulnerable to active attacks due to usability problems and incomplete mental models of public key cryptography [15]. This study included 28 computer scientists; of the participants, four clicked through the warning message, eight could not find the ceremony, and ultimately only seven were able to successfully authenticate their peer. Assal et al. asked participants to perform the authentication ceremony in ChatSecure using different key representations, which include a fingerprint, shared secret, and QR code [1]. Of the 20 participants in this study, 20% were successful for the fingerprint, 85% for the shared secret, and 30% for the QR code.

Two papers have compared the usability of various fingerprint representations. Tan et al. compared eight representations, including textual and graphical representations with varying degrees of structure, in a simulated attack scenario [18]. Graphical representations were relatively more susceptible to attack, but were easy to use and comparison was fast. Participants used different strategies for comparison, often comparing only a portion of the fingerprint or comparing multiple blocks at a time. Dechand et al. studied textual key verification methods, finding that users are more resistant to attacks when using sentence-based encoding as compared to hexadecimal, alphanumeric, or numeric representations [6]. Sentence-based encoding rated high on usability but low on trustworthiness.

Herzberg and Leibowitz examined the usability of WhatsApp, Viber, Telegram, and Signal, finding that most users were unable to properly authenticate, both in an initial authentication ceremony and after a key reset [8]. The study included 39 participants from a variety of backgrounds and all were given instruction on end-to-end encryption. Most users failed to authenticate on the first attempt; they were then given additional instruction about authentication. About three-quarters authenticated properly after the additional instruction was given.

Our work differs from these studies in several important ways. First, we study in detail the ability of users to discover and use the authentication ceremony in a variety of secure messaging applications, giving us insight into the differences among these applications. Schröder et al. only study Signal, and Dechand et al. do not study any particular applications. Second, we use a paired participant methodology, so that users are asked to identify a friend they already know, rather than an unknown study coordinator. This method is more realistic than most prior studies and yields important insights into user behavior. For example, our study participants called each other, verified through voice and vision, and asked questions based on shared knowledge. Third, we conduct a between-subjects study on the effects of instruction, so that those receiving instruction are not biased by their previous experiences. The first set of participants were asked to authenticate given only general awareness of threats, while the second set of participants received instruction about the importance of comparing encryption keys.

Another important aspect of our work is that it provides replicability that is not possible with prior work. Herzberg and Leibowitz report a similar result, that participants authenticated properly after additional instruction about authentication was given. However, their paper provides few details about the instruction given and does not report detailed statistics, so it is difficult to draw any quantitative conclusions about the effect of the instruction or the relative merits of the different applications they tested. We report detailed statistics about what methods users tried with each application, the time taken to authenticate, SUS scores, trust ratings, and favorite systems. We include our full study materials in the appendix and provide our dataset on a companion web site.

Significant work in the area of secure email has also examined issues related to usable authentication. Obtaining and verifying the key for a recipient is an important use case for email, and lessons learned may apply to secure messaging as well. Numerous papers attest to the difficulties users have with this and other key management steps [23, 16, 12].

The most success in this area has been in the use of automatic authentication using a trusted key server. Bai et al. [3] has shown that individuals recognize the security benefits of manual key exchange, but prefer a centralized key server that authenticates users and distributes keys associated with their email address, due to greater usability and "good enough" security. This model has been simulated by Atwater et al. [2] and implemented using IBE by Ruoti et al. [11]. Likewise, the use of secure messaging applications is generally considered a success for automatic key management.

3. APPLICATION DESCRIPTIONS

The three secure messaging applications used in our study are WhatsApp, Viber, and Facebook Messenger. These three applications were chosen because they present users with distinct key verification experiences and because of their popularity and large installation base.

3.1 WhatsApp

WhatsApp is perhaps the most well-known and widely-used messaging application, boasting a user base of over one billion users. While it did not originally offer secure messaging functionality at its inception, in November of 2014, WhatsApp partnered with Open Whisper Systems to incorporate end-to-end encryption using the Signal encryption protocol.

When a conversation is initiated, WhatsApp inserts a message informing users that messages they send are encrypted with end-to-end encryption. Users are given two options for key verification: QR code scanning and key fingerprint verification (both parties see the same fingerprint). In accessing this dialog, a short caption accompanies the "Encryption" option in the previous menu, informing users that they can "Tap to verify." Doing so brings up the verification dialog shown in Figure 1a.

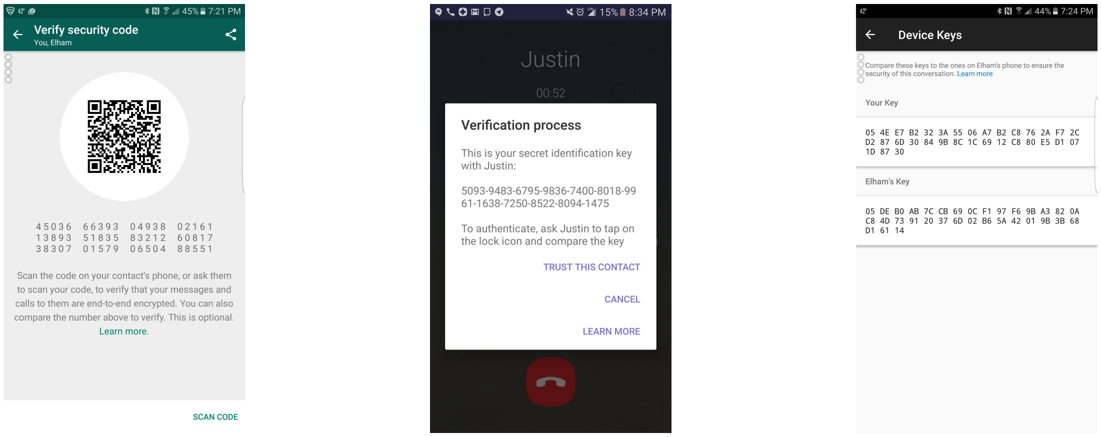
3.2 Viber

Viber is another widely-used messaging application with an install base of over 800 million users. As with WhatsApp, it did not originally offer end-to-end encryption, adding this functionality in April of 2016. Its encryption protocol is a proprietary design allegedly based on the principles of the Signal protocol.

While—as with the other two applications—Viber does not immediately make apparent the need to verify keys, once begun, it does—unlike the other two applications—carefully guide the user through the process with a set of instructional dialogs. In displaying this functionality, Viber does not opt to use the terms "encryption" or "key" at the outset, instead characterizing the verification process as "trust[ing]" one's conversation partner. Only after the user selects this option, are they prompted with a dialog that explains the need to confirm that "secret keys are identical." This process is facilitated via a free Viber call. After making the call, both sides may see their keys by tapping a lock icon that appears during the call, allowing for verification. This dialog is shown in Figure 1b. It should be noted, however, that Viber does not allow the user to view their keys without initiating this call, nor does it allow the user to view these keys once a contact has been marked as trusted.

3.3 Facebook Messenger

Facebook Messenger is the messaging utility designed by Facebook to integrate into their chat system, and, like WhatsApp, has a user base of over 1 billion users. Again, as with the other two applications, it did not originally offer end-to-end encryption, adding this functionality in October of 2016. It also uses the Signal protocol.



(a) WhatsApp (b) Viber (c) Facebook Messenger

Figure 1: Authentication ceremonies in each of the applications.

The user experience of Facebook Messenger's encryption functionality differs substantially from WhatsApp and Viber. While the first two applications encrypt all communication automatically, Facebook Messenger defaults to an unencrypted chat session, with users required to initiate a standard chat session before accessing a "Secret Conversation" function via the conversation menu. Once within the secret conversation menu, users can access their device keys via the context menu. At this point, the experience again diverges from the two other applications, as the key verification dialog presents users with two keys instead of one. Furthermore, the Facebook Messenger key verification interface does not easily facilitate a way for users to communicate these key values to the other party. This dialog is shown in Figure 1c.

4. METHODOLOGY

We conducted an IRB-approved, two-phase user study examining how participant pairs locate and complete the authentication ceremony in three secure messaging applications: WhatsApp, Viber, and Facebook Messenger. Our study materials are shown in Appendix B and our full data set is available at `https://alice.internet.byu.edu`.

In both phases, we asked participants to complete a scenario where one participant needed to send a credit card number to the other participant. We instructed participants to verify that they were truly communicating with their partner and to ensure that no other party could read their messages. Our instructions informed participants that the application would help them accomplish these goals, but they were left in control of the methods used to ensure these conditions were met. In the second phase, participants viewed and read aloud an instructional set of slides that informed them about the importance of comparing encryption keys.

Each phase was a within-subjects study, and all participants used all three applications in each phase. The participants differed between the two phases, allowing us to see between-subjects differences in instruction between the two phases.

To choose the three applications we compared the authentication ceremony in 10 secure messaging applications—WhatsApp, Telegram, Signal, Zendo, Facebook Messenger, Viber, ChatSecure, Allo, Line, SafeSlinger. We binned the applications into groups, based on the authentication methods used. We then narrowed our choices to the following: Signal/WhatsApp (use both QR codes and manual verification), Telegram/Facebook Messenger (use manual verification, include non-secure chatting), and Zendo (uses NFC or QR code, requires verification before chatting). We chose WhatsApp over Signal and Facebook Messenger over Telegram because of their greater popularity in the United States. As explained below, we were unable to proceed with Zendo in the study. We chose Viber as an alternative because it provides a method for manually comparing encryption keys using a phone call built into the application. This provided us with three different applications that use a variety of authentication methods.

4.1 Pilot study

We conducted a pilot study of the first phase with three pairs of participants, using WhatsApp, Facebook Messenger, and Zendo. The Zendo secure messenger employs key verification as a forcing function: users must first scan each other's QR codes, or use NFC communication, before the conversation can begin. Unfortunately, we experienced multiple, severe technical difficulties with the application during the pilot study, leading us to abandon it in favor of Viber.

4.2 Study recruitment and design

We placed flyers advertising the study around the campus of a local university. These flyers contained a link that participants could use to schedule online, and they included a requirement that all participants bring a friend and smartphones in order to take part in the study. Recruitment proceeded from February 3, 2017 to February 28, 2017, with 39 unique participant pairs being recruited in total: 12 for the first phase of the study, and 24 for the second.[1]

[1]One second-phase participant pair experienced difficulty because one participant had limited English proficiency and our study was executed entirely in English (this participant thought that they were being tasked with locating a physical key). Technical errors occurred during the data collection of two other pairs and they were presented with incorrect post-task questionnaires. Accordingly, the data for these three pairs were excluded from the study and we recruited replacements in their place.

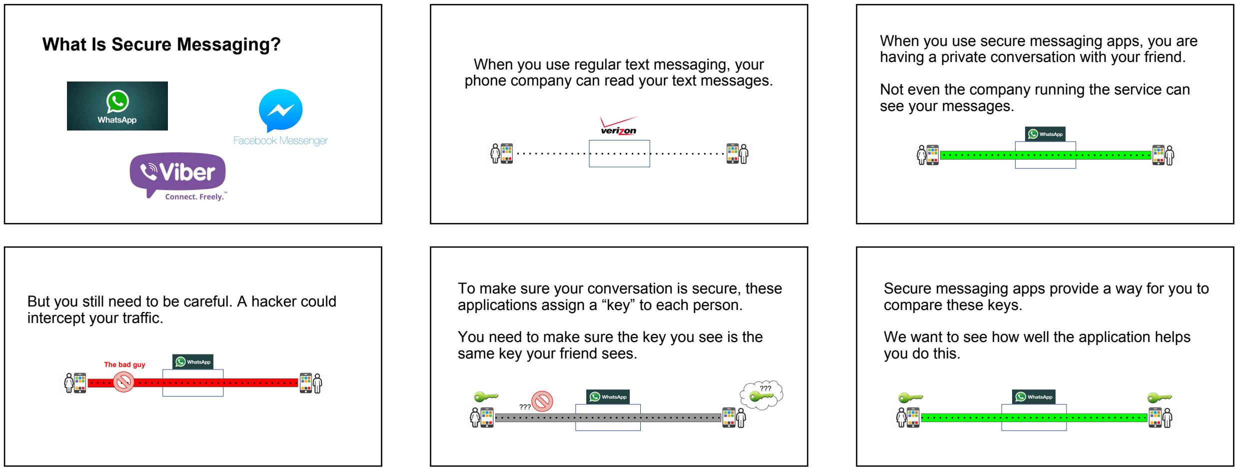


Figure 2: Instructional slides used in the second phase.

To ensure different pairs of participants tried applications in different orders, we calculated a complete set of permutations listing the order in which each of the three applications would be used by a given pair. We then randomized the permutation that was assigned to each participant. This ensured a collectively uniform distribution of sequences while keeping the assignment of a given sequence to a particular pair random. Each ordering of the three systems occurred exactly twice in the first phase and four times in the second.

The study was conducted in two phases, spanning a period of one month. The first phase ran from February 3, 2017 to February 16, 2017. It took roughly 40 to 45 minutes for each pair of participants to complete, for which they were compensated $10 USD each. The second phase ran from February 17, 2017 to March 2, 2017. The second phase studies were more involved and took longer to complete, roughly 60 minutes each, and so all participants were compensated at a higher rate of $15 USD.

When participants arrived for their scheduled appointment, we presented them with the requisite forms for consent and compensation. We instructed them to download and install any of the three applications—WhatsApp, Viber, and Facebook Messenger—that they did not already have on their phones, to minimize the likelihood of technical difficulties during the study.[2] We then read them a brief introduction describing the study conditions and their rights as study participants. We informed them that they would be placed in separate rooms, but could freely communicate or meet with one another if they deemed it necessary to complete their task. We also informed participants that a study coordinator would be with them at all times and would answer any questions they might have.

We randomly assigned one member of each pair as Participant A, with his or her counterpart becoming Participant B, delineating their roles in the subsequent tasks. We then led them to their respective rooms, seating them at a computer, and initiating audio recording. We preloaded each computer with a Qualtrics survey that guided participants through the study, and it included a demographic questionnaire, instructions regarding the three tasks they were to perform, and post-task questionnaires. Each of the three tasks was identical in nature, differing only by which of the three secure messaging applications participants were to use to complete the task. Throughout the study, study coordinators were available to answer general questions about the study, but were careful not to provide any specific instructions that would aid in the use of the applications themselves.

4.3 Task design

In both phases, the tasks participants completed were the same: Participant A was to securely retrieve a credit card number in Participant B's possession by using the application that was being tested. This scenario was intended solely as a narrative backdrop for the tasks we were truly concerned with: finding and completing the authentication ceremony. When asked to complete the task, participants were instructed as follows:

> *Your task is to make sure that you are really talking to your friend and that nobody else (such as the service provider) can read your text messages. The application should have ways to help you do this.*

Accordingly, despite a difference in roles, there were no practical differences between the tasks Participant A and Participant B needed to complete. Participants were instructed and encouraged to "talk aloud" as they completed the task, explaining the choices they made and the actions they took.

Additional instruction was given in the second phase. Before participants were introduced to the task, they were asked to read aloud a short set of slides, shown in Figure 2. These slides informed them that traffic interception was a possibility, that secure messaging applications accordingly provide a "key" that could be compared to ensure that conversations were indeed secure, and that they needed to make sure that they saw the same key as their counterpart. Furthermore, on the

[2]In our pilot study, several participants lacked sufficient space on their phones to install the applications or had phones that were too old to run the applications properly. We subsequently adopted this measure in an attempt to forestall re-occurrence.

first task, if second phase participants had failed to verify one another's identity either prior to sensitive data exchange or after ten minutes had passed, they were marked as having failed the task and prompted by study coordinators to look for a way to authenticate properly.

4.4 Study questionnaire

Participants were led through the study by a web-based Qualtrics survey. We first discuss those aspects that were held constant for both phases, followed by an explanation of how the questionnaire differed in the second phase.

Upon beginning the survey, participants first answered a set of demographic questions. They then answered questions about their past experience, if any, with secure messaging applications. This included questions about which applications they might have used, their reasons for doing so, and their general experiences with sending sensitive information. Participants were next shown a description of their first task (all three tasks were identical, diverging only on the system being used). Each task was followed with a post-task questionnaire assessing their level of trust in the application, whether or not they believed they had successfully verified their partner's identity and why, and who they believed was capable of reading their conversation. After all three tasks had been completed, participants were then asked which of the three applications was their favorite and why.

In the second phase, participants were given supplementary instructions and asked additional questions. First, after the demographic questions, participants were asked a series of six questions intended to gauge their relative familiarity with end-to-end encryption. Next, prior to beginning the first task, they were presented with, and asked to read aloud, a set of six slides that very briefly introduced the role of keys and informed them that the applications they were about to use would provide a way for them to compare these keys. These instructional slides are shown in the appendix. Finally, at the end of each task, the post-task questionnaire from the first phase was augmented by the ten questions from the System Usability Scale (SUS).

4.5 Post-study debrief

At the conclusion of each study, participant pairs were gathered in the same room and asked a series of questions. This served as a complement to the questionnaires that they had answered individually, and gave them an opportunity to react to one another. Participants were prompted regarding incidents specific to their experience—e.g., if they had evidenced visible frustration with a particular app—as well as general questions. Examples of the latter include having participants contrast the authentication ceremony used by each application, as well as asking them to explain what role they thought keys played in verifying one another's identity.

4.6 Demographics

Our sample population skewed slightly female (n=40, 56%) and young, with 74% (n=53) between the ages of 18 and 24, and 26% (n=19) between 25 and 34. Because we distributed recruitment flyers on a university campus, most of our participants were college students (n=48, 61%), with 17% (n=12) having less educational experience than that, and 22% (n=16) having at least finished college. Participants had a variety of backgrounds, with roughly even representation between technical (i.e., STEM; n=34, 48%) and non-technical backgrounds (n=37, 52%), and 10 (14%) in explicitly IT-related fields. (One participant failed to identify their field of study or occupation.)

In the second phase, the questionnaire included a series of six multiple-choice questions intended to assess participants' knowledge of end-to-end encryption. We assigned equal weights of one point to each question, and scored each participant from 0-6, corresponding to the number of correct answers given by the participant. Participants were further placed into categories of "beginner," "intermediate," and "advanced" for scores in the range of 0-2 for beginners, 3-4 for intermediate, and 5-6 for advanced. There were an equal number of participants with beginner and intermediate ratings—21—with 6 participants netting an advanced rating. Beginners were mostly female (3:18), intermediate participants were mostly male (15:6), while the advanced category had an even gender split (3:3).

4.7 Limitations

The instructions given to the first three participant pairs of the first phase were slightly different from those given to the remaining nine. They were directed to ensure that no one was "listening in" on their conversation, a directive participants took literally as they would visibly scan the room for potentially intrusive parties. This wording was subsequently altered, with participants instead instructed to ensure that "nobody else (such as the service provider) can read your text messages."

The slides we provided participants to teach about cryptographic keys were necessarily simplified so that they could be understood by novices. In this material we mentioned that participants should ensure the key they see is the same as their partner's. While this was sufficient in describing tasks for Viber and WhatsApp, Facebook Messenger actually utilizes two keys, one for each partner. This subtlety was not mentioned by any participant nor did it seem to adversely affect their performance.

Finally, due to our method of recruitment, our participants were largely students and their acquaintances, and subsequently exhibited some degree of homogeneity, e.g., all participants were between 18 and 34 years of age. They are thus not representative of a larger population. Furthermore, while an effort was made to place participants in a more organic setting—e.g., by having them communicate with real members of their social circle as opposed to study coordinators—this was still ultimately a lab study and has limitations common to all studies run in a trusted environment [10, 17].

5. FIRST PHASE RESULTS

In the first phase of the study, only 2 of the 12 pairs experienced some success in locating and completing the authentication ceremony, with an overall success rate of 14% across all pairs and applications.

Participants used a variety of ad hoc methods for authentication. Listed in the order they appear in Table 1, these methods were: utilization of a picture for visual identification, utilization of a live video feed for visual identification, utilization of shared secrets for identification, utilization of contact information (e.g., phone number, profile picture) for identification, utilization of a shared second language for

Application	Send Picture	Recognize Video	Recognize Voice	Shared Knowledge	Contact Info	Second Language	Authentication Ceremony
WhatsApp	0	0	13	10	3	2	2
Viber	0	10	4	7	2	2	4
Facebook Messenger	2	12	2	7	0	0	2

Table 1: Methods of authentication used in the first phase by pairs of participants.

identification, and performing the actual authentication ceremony. These categories were compiled by asking users how they authenticated the other party, and are not mutually exclusive (some used more than one method).

We examined the two pairs that were successful to better understand their experiences. One pair was successful because of their curiosity, which led to them exploring the application settings. This pair started with Viber and began to verify each other simply through a phone call, when they suddenly noticed the option in Viber to authenticate a contact, making that contact "trusted." They subsequently verified the encryption key through the phone feature embedded in the authentication ceremony. After this experience, this pair noticed they should be looking for similar functionality in the other applications. The followed the on-screen instructions in WhatsApp to scan the QR code, and they exchanged a screenshot of the authentication code in Facebook Messenger.

A second pair started the study with Facebook Messenger. This pair called each other using an insecure phone call, spoke in Korean, and transferred the credit card number used in the scenario without completing the authentication ceremony. They next used WhatsApp, and because it was their first time using the application, they were prompted with a notice about end-to-end encryption after sending their first message. After clicking to learn more, this pair was able to locate and complete the authentication ceremony by using a phone call to read and verify the key. After this experience, the pair was also able to locate the lock icon in Viber, follow the instructions in the ceremony, and use a phone call to verify the key. However, they were unsure about the role of the key and still verified each others' identity by asking questions that relied on their common knowledge.

6. SECOND PHASE RESULTS

In this section we discuss results regarding participant use of the authentication ceremony for the second phase, when additional instruction was given regarding the importance of comparing keys.

6.1 Success Rate

The success rate for completing the authentication ceremony in the second phase was drastically higher than for the first phase. Overall, the success rate was 78% across all participant pairs and the three applications. Table 2 shows the breakdown of the success rate for each application. Failures occurred when participants transmitted sensitive data before verifying keys, or if they failed to find and validate the keys within ten minutes of opening the application. Successes indicate that participants identified and compared keys in some fashion. The Error column indicates three cases where Facebook Messenger failed to deliver messages or failed to display important UI elements that allow participants to access key information. We noted various mistakes made by participants, but these were considered distinct from failures and are discussed later.

Application	Success	Fail	Error
WhatsApp	19 (79%)	5 (20%)	0 (0%)
Viber	23 (96%)	1 (4%)	0 (0%)
Facebook Messenger	15 (63%)	6 (25%)	3 (13%)

Table 2: Success rates per pair of participants for the authentication ceremony in the second phase.

The leap from a 14% success rate in the first phase to 78% in the second phase suggests that users are capable of locating and performing the authentication ceremony when prompted. Some of these applications indicate that keys need to be validated, yet our results from phase one indicate that these instructions are largely ignored, thus we suspect that the independent prompts from our study accounted for much of the difference seen in authentication ceremony success rates.

To test whether there are any differences between the applications, we used Cochran's Q test. We found that the success rate was statistically different for the applications ($\chi^2(2) = 15.429$, $p < .0005$). We then ran McNemar's test to identify the significant differences among the pairs of applications. We found there is a significant difference between WhatsApp and Viber ($p = 0.008$) as well as between Facebook Messenger and Viber ($p < 0.0005$).

It is interesting that Viber's success rate is significantly higher than the other two applications. Viber's authentication ceremony uses an in-app phone call and provides a UI that helps users view and read the encryption key during the phone call. Both Facebook Messenger's authentication also provides only manual verification of the encryption key, but does not provide this assistance.

6.2 Verification Methods

The methods used by participants to perform the authentication ceremony are shown in Table 3. Note that some participants used more than one method. We do not include methods for three pairs of participants who encountered errors when utilizing Facebook Messenger. These errors prohibited us from assessing how these participants would have interacted with the authentication ceremony.

The most-selected method for the ceremony through WhatsApp was scanning the QR code of the key fingerprint in person. Of the applications we studied, this method is unique to WhatsApp. Some pairs opted to take a screenshot of the key or QR code and send it this way, while others remembered substrings of the key fingerprint and repeatedly visited the text screen to send pieces of it to their partner. This behavior

Action	WhatsApp	Viber	Messenger
Secure Methods			
Scanned QR code in person	11 (46%)	N/A	N/A
Read key in person	1 (4%)	0 (0%)	7 (29%)
Called out of band or used Viber's call method to provide key	1 (4%)	23 (96%)	1 (4%)
Less Secure Methods			
Sent key through in-app text	7 (29%)	N/A	10 (42%)
Sent key through in-app video	3 (13%)	N/A	4 (17%)
Sent key through in-app voice	1 (4%)	N/A	1 (4%)
Failures			
Sent sensitive information before validation	5 (21%)	1 (4%)	5 (21%)
Failed to find key within 10 minutes and after a hint	1 (4%)	0 (0%)	1 (4%)

Table 3: Methods used for the authentication ceremony in the second phase. Numbers indicate pairs and percentages are out of the total number of pairs.

occurred when participants discovered the QR code and key fingerprint but were confused as to what to do next.

Numerous participants using WhatsApp read the key data in person, read the key using a voice or video call, or sent the key using text. Most participants using Facebook Messenger used these methods, since they were the only ones available.

Viber provides a much stricter interface once a user has located the option to verify his partner's identity. Instead of offering key material immediately, an in-app call must be initiated before the key material is provided to the user. As a result, all pairs who successfully completed the ceremony utilized this feature to verify their keys. We note that this policy resulted in no mistakes made for the authentication ceremony. However, the process confused some participants, and three pairs sent sensitive information through the application without performing this procedure.

6.3 Timing

We timed each pair of participants to obtain two metrics: the time taken to locate and identify the authentication ceremony as it is presented within the application interface and the time taken to complete the ceremony successfully. In the case of finding the ceremony, the time reported is the time taken for the first partner to identify the key material or complete the task. We consider timing data only for cases where the pair succeeded in authenticating successfully because we stopped participants after 10 minutes if they could not find the ceremony.

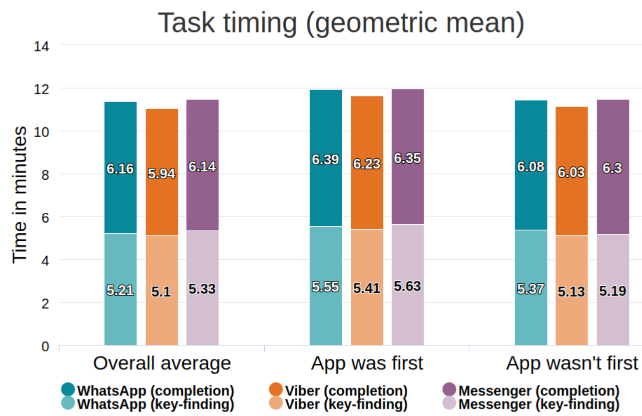


Figure 3: Timing for finding and using the authentication ceremony in the second phase. Lighter shades indicate the time taken to find the ceremony and the full bar indicates time taken for completing the ceremony.

Figure 3 shows the geometric mean of both time metrics for the three applications tested.[3] Applications that are selected to be evaluated first in a given study have a disadvantage with respect to time because it is users' first exposure to the task and possibly keys in general. To account for this, Figure 3 also includes comparisons showing timing data from when each application was studied first and when the application was not studied first.

To test whether there is a significant difference in the time to complete these tasks among the three different applications, we used the Kruskal-Wallis test. We found that there are statistically significant differences among the applications for both finding the ceremony ($p = 0.031$) and completing the ceremony ($p = 0.043$). We next ran pairwise post-hoc Dunn's tests to determine where the differences occur. We found a significant difference between Facebook Messenger and WhatsApp for finding the ceremony ($p = 0.030$), with Facebook Messenger being faster (mean time, Facebook Messenger=2.5 minutes, WhatsApp=3.7 minutes). We also found a significant difference between Viber and WhatsApp for completing the ceremony ($p = 0.045$), with Viber being faster (mean time, Viber=6.9 minutes, WhatsApp=8.5 minutes).

A major takeaway from the timing data shown is that key discovery and key verification both require substantial time for all three applications. On average, across all applications discovery of the ceremony required 3.2 minutes and ceremony completion required another 7.8 minutes. Given that the participants were informed about the existence of the keys beforehand and told explicitly to verify them, these times are unsatisfactory from a usability standpoint. The usability issues and concerns voiced by participants responsible for these times are discussed in Section 7.

7. APPLICATION FEEDBACK

In this section we discuss feedback that participants provided regarding the secure messaging applications, including usability, their favorite application, and the trustworthiness of the applications.

[3]Sauro and Lewis recommend using the geometric mean for task timing [14] because timing data is positively skewed and the geometric mean reduces error.

SUS subcategory	WhatsApp	Viber	Messenger
Overall	65.45	67.45	**67.78**
First system	65.47	67.97	**69.22**
Not first system	64.45	66.02	**67.97**
Success	64.41	67.86	**72.71**
Failure	66.25	63.13	**69.50**

Table 4: SUS scores for the applications in the second phase.

7.1 Usability

During the second phase of our study, participants evaluated each application using the System Usability Scale (SUS). Table 4 presents the breakdown of the scores for each system across various subcategories. The values shown are the mean values for each subcategory, while bolded values highlight the highest SUS score for each subcategory.

We report SUS scores across five subcategories for each application: overall SUS score, the mean SUS score when the application was the first of the three presented, the mean SUS score when the application was not the first shown, the mean SUS score for participants who succeeded at the task using the given application, and the mean SUS score for participants who failed the task.

Although SUS scores range from 0 to 100, this is not a percentile value and can thus be difficult to interpret. Accordingly, to help contextualize the values shown, we draw on the findings of researchers familiar with SUS. Sauro [13], extending work from other researchers such as Bangor et al. [4], created a guide for interpreting a given SUS score by normalizing it relative to those achieved by other systems. This framework associates SUS scores with percentile rankings and with letter grades (from A+ to F).

For reference, the applications' overall SUS scores fall within the "C" range, landing somewhere within the 41st to 59th percentile. The single lowest SUS score—Viber's mean failure score—nets a "C-" grade, falling within the 35th to 40th percentile. The highest SUS score—Facebook Messenger's mean success score—achieves a "C+" grade, somewhere within the 60th to 64th percentile.

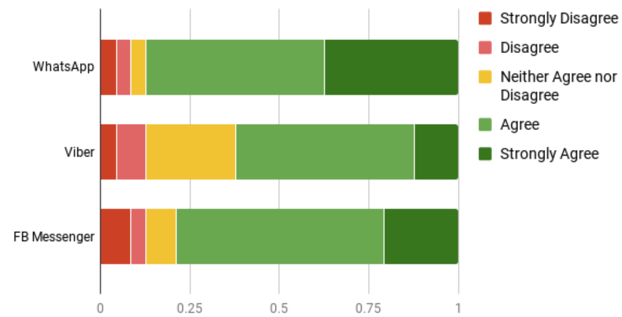
7.2 Favorite application

Participants were asked to select which, if any, of the three applications was their favorite and why. Table 5 shows the breakdown of responses for each phase. Facebook Messenger was the most preferred system, followed by WhatsApp. We ran a Chi-Square test to determine if the differences in the ratings between phase one and phase two were statistically significant and they were not.

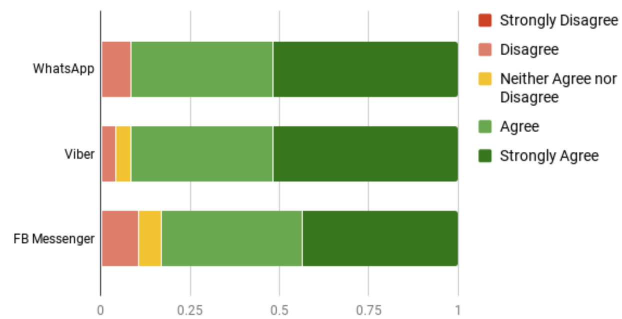
Though numerous reasons were given for why a particular system was a participant's favorite, familiarity was by far the most commonly cited reason for preference (except with Viber, which was not previously used by any of our participants). The next most common reason given, and one that held true for each of the three systems, was ease-of-use, with what constituted "easy to use" varying from system to system. Some WhatsApp users, for example, appreciated its ability to scan QR codes for key verification, obviating the need to read aloud the long string of digits comprising a key fingerprint. Those who liked Viber found its key verification

Study phase	WhatsApp	Viber	Messenger	None
One	39.1%	8.7%	47.8%	4.4%
Two	31.3%	22.9%	43.8%	2.0%

Table 5: Participants' favorite applications. Each cell contains the fraction of participants from each phase who, when prompted for their favorite system, gave the respective response.



(a) Trust ratings in the first phase.



(b) Trust ratings in the second phase.

Figure 4: Participant ratings of trust for each application.

process the simplest to access and execute. By contrast, those who mentioned ease-of-use relative to Facebook Messenger typically associated it with familiarity as opposed to any mechanism in particular.

7.3 Trust

As part of each post-task questionnaire, participants were asked to rate their trust in each application. They were presented with the statement *"I trust this application to be secure"* and asked to rate the statement on a 5-point Likert scale ranging from "strongly disagree" to "strongly agree." Responses for the two phases are shown, normalized, in Figure 4.

Comparing the trust scores from the two phases, two points stand out. First, a "strongly disagree" response—indicating a total lack of trust in the application—appeared for all three of the applications in the first phase, but not at all in the second phase. This is mostly due to one participant from the first phase who chose "strongly disagree" for all three systems. Secondly, responses of "strongly agree"—indicating confidence and trust in the application—are much more prevalent in the second phase.

To compare the trust scores in more detail, we ran a mixed model ANOVA Test, which allowed us to see the the interaction between the two independent variables (application and phase). We found that there is a significant interaction between the application and the study phase (F(2,140) = 5.023, $p = 0.008$, partial $\eta^2 = 0.067$).

To determine whether there was a simple main effect for the application, we ran a repeated measures ANOVA on each phase. There was a statistically significant effect of the application on trust for phase one (F(2,46) = 4.173, $p = 0.022$, partial $\eta^2 = .154$). By examining the pairwise comparisons, we found that the trust score was significantly lower for Viber as compared to WhatsApp in the first phase (M = 0.542, SE = 0.180, $p = 0.19$).

To determine whether there was a simple main effect for the study phase, we ran a one-way ANOVA on each application to compare the trust between the two phases. There was a statistically significant difference in trust ratings between the two phases for Viber (F(1,70)=14.994, $p < 0.0005$, partial $\eta^2 = .176$). The mean trust for Viber in the first phase was 3.58, and in the second phase it increased to 4.40.

Altogether, this analysis indicates that Viber was trusted less than WhatsApp in the first phase, but then was trusted significantly more in the second phase, after some instruction about the importance of the authentication ceremony. The trust for Viber increased in the second phase to the point that it was not significantly different from WhatsApp.

Participant commentary raised two other points of interest. First, participants strongly associated reputation with the trustworthiness of applications. Viber, for example, despite possessing a large user base outside of the United States, was essentially unknown to our participants, leading them to express wariness of this application. Facebook's status as a household name both inspired confidence and distrust. While its reputation as a large and established company reassured some, others were discomfited by the many negative stories they had heard about account hacks and privacy invasions on Facebook. Second, responding to descriptions of end-to-end encryption and promises of secure communication by the various applications, multiple participants remarked that they had no way to truly gauge the validity of those statements. Both these sentiments are captured by a remark from R10B, *"I would say it's a double-edged sword because Facebook—everyone knows Facebook—but it has that reputation of getting hacked all the time. But I've never heard of Viber or WhatsApp, so it could easily be some third-party Ukrainian mean people who want to steal information because that's just who they are. And whether it states that they're not gonna read or listen to the conversations and stuff like that... well, who knows?"* However, most opted to believe, for as one participant concluded, *"at some point, you have to trust **something**."*

8. OBSERVATIONS

During our study, certain participant experiences and commentary stood out, highlighting a handful of concerns about each of the three applications individually, and in general. We feel that these observations are worthy of note in that they suggest directions for focus and improvement in the domain of secure messaging.

8.1 Single key fingerprint

WhatsApp and Viber both generate a single key fingerprint to be shared between pairs. While alternating recitation of segments of the key is likely the intention of developers, in practice, relationship dynamics complicate the issue. We observed several instances where the dominant partner in the relationship read the entire key on their own, with their partner simply offering an affirmation in response. When key verification is done in this manner, one party never actually demonstrates any knowledge of the shared secret—it is entirely possible that a man-in-the-middle could simply convey validation of the key when their key is, in actuality, different. This effect is further emphasized when, as we saw in one instance, the listening party asks the speaking party to repeat the first part of the key, reinforcing the speaking party's belief that their partner is in possession of the correct key. It is, however, worth noting that this "extended" validation once again did not demonstrate any actual knowledge of the secret.

8.2 Key length

It was often observed during the study that participants were surprised at the length of the key data they were intended to relay to their partners. Though every application used a form of fingerprinting to greatly reduce the total characters that needed to be read, users often verbally remarked that strings were too long. During the key exchange process we often witnessed fatigue, as participants would read only half the key and claim that was "good enough" and some recipients even ignored the key being read to them by their partners after the first few numbers matched. R27A used a QR code transmission to handle her first authentication ceremony with WhatsApp. Upon realizing that no such option existed for Viber, her second application used, she looked at the key and exclaimed, *"It's about eight years long!"*. R27A successfully checked every digit of the key data with her partner, but voiced her disapproval of its length repeatedly throughout.

8.3 Viber-specific issues

We observed two issues with Viber. The first relates to its mechanism for verifying a new user's phone number. While most applications send a confirmation text containing a code, as does Viber, it nevertheless defaults to calling the new user first as a primary and alternative confirmation mechanism. This took many of our participants by surprise and left them ill-at-ease to see an unknown number suddenly calling them. Secondly, and far more concerningly, Viber does not provide a mechanism to revoke trust. While this is likely a conscious decision on the developers' part, it can cause issues in practice. More specifically, one participant inadvertently tapped the trust button while trying to figure out how to verify his partner's key, thus accidentally conveying to the application in an apparently irreversible manner that this individual was now trusted.

Many users were also critical of the Viber UI's phrasing for the option to begin the process of key verification. The option is labeled "Trust this contact," which many users hesitated to press, unsure if it would inform the application to trust the contact or if it would bring up further dialogues to perform the validation. R36A visibly hesitated during this step during the study and articulated this concern in the exit interview: *"if I click 'Trust this Contact' but I haven't verified [my partner] yet, it's kind of weird."*

8.4 WhatsApp-specific issues

We observed several issues with WhatsApp. WhatsApp appends a pair of checkmarks next to each message, representing the delivery and read status of the respective message. However, a handful of participants mistakenly associated these checkmarks with security, operating under the misconception that a checkmark beside a given message indicated that it had been secured. The other two issues concern the key verification mechanism. When a matching QR code is scanned, the application briefly flashes a green checkmark logo over the QR code area, indicating that the fingerprint has been validated and is correct. However, because it disappears quickly, leaving no lasting indication that verification has occurred, numerous participants wondered if they had verified the key or not. Additionally, the key verification screen includes a button to share a screenshot of the verification screen. Some of our participants assumed that they could use this to send a screenshot to their partner, who could then scan the QR code contained therein. Unfortunately for them, WhatsApp does not provide functionality to scan a QR code from an image, serving to confuse those who tried.

8.5 Facebook Messenger-specific issues

In addition to the usability concerns already described, such as the difficulty in locating device keys, Facebook Messenger's Secret Conversation functionality—its mechanism for secure communication—errored more than a few times during our study. More importantly, however, was that these errors were not apparent to participants. Participants were thus unaware that the Secret Conversation was not operating as intended, and instead blamed themselves or their counterparts for failure. One example we encountered several times was that encrypted messages sent via this mechanism appeared normally on the user's phone despite never being received by their partner. One such participant began shouting in exasperation at her phone, exclaiming, *"I feel like I am having a conversation with myself! What's wrong with this app?!"*

8.6 Key changes

One important issue that secure messengers must deal with in practice is a key change occurring mid-conversation. As this was not tested by our participants during our study, we recreated this scenario in each of the three applications to observe their respective reactions. Facebook Messenger inlines a message when one's conversation partner's key changes, informing the user that their device has changed and that their key has changed. While it does not explicitly instruct the user to re-verify the key, of the three applications, it makes the user aware that key change has occurred. Viber gives no proactive notification to the user that key change has occurred, but when the conversation menu is again accessed post-change, Viber includes an explicit message warning the user that they will need to re-verify the identity of their conversation partner. WhatsApp presented no notification that we could observe. It neither inlined a notification as Facebook Messenger did, nor does it indicate to the user that re-verification must be performed. In fact, WhatsApp presents no lasting UI change that allows a user to confirm that verification has occurred at all.

9. USER THREAT MODEL

Two authors jointly coded responses to two survey questions used in both phases regarding participant perception of the authentication ceremony. These questions were:

- Please explain why you think you have (or have not) verified the identity of your friend.
- Who do you think can read your message except you and your friend?

In reviewing the coded data, some details of the threat models perceived by users became evident.

Note that, if correctly followed, completing the authentication ceremony successfully guarantees that a participant has authenticated their partner and no other party can listen in on the conversation. This of course assumes that the applications have properly implemented cryptographic protocols. None of the applications studied are open source, so their claims cannot be verified.

Of the 141 times the first of these prompts was presented (excluding Facebook Messenger errors), 109 responses indicated that the authentication ceremony was a primary reason for successful identification. This is encouraging, but also expected given the focus that the study placed on its significance, which may have biased participants. For example, in response to the first prompt, R13B stated *"...I asked him a person[al question] that he responded [to] in the right manner, but also because our messages were encrypted and our personal keys matched."* The use of questions that rely on shared knowledge was a common response to this prompt, and it was often coupled with a reference to verifying the key.

Where verification of personal inquiries are mentioned in tandem with key verification as a reason for verified identities, it is unclear whether participants believe the inquiry can be used as a substitute for key verification or if they are expressing the more secure notion that proper key verification includes explicit identity matching. To mitigate any mislabeling due to this lack of clarity, we focus on the responses that did not mention key verification as the reason for identity verification, which occurred 32 times. These responses focused on verifying features of their partner and considered impersonation or physical duress attack vectors. For example, R24A asserted he had verified the identity of his partner because he had *"asked personal questions that are difficult to know from online material/searches* and R36B confided that his partner *"was able to tell [him] something that no one else would know. Unless he was being held at gunpoint."* Of these 32 responses, 28 (88%) of them mention using features of their partner as the method of verifying identity (e.g. physical appearance in video, shared private knowledge, familiar voice). Two others mentioned trust in the application itself, one admitted no attempt to verify, and one trusted that their partner verified on their behalf.

The second prompt listed above provided some insight into the set of possible attackers considered by participants. This question was issued 141 times as well, immediately following the prompt mentioned earlier. Though 109 responses indicated that the identity of their partner had been verified, only 76 (70%) responses indicated that no other party could read messages exchanged between the two partners. The responses of those who indicated that other parties may be privy to the information were coded to determine the nature of the suspect parties. Five distinct entities were found to be mentioned in those responses: government, cellular service providers, physical accessors (e.g., shoulder surfers,

Type	Times Mentioned
Service Provider	4
Government	8
Hackers	17
Physical Accessors	18
Application Developer	19

Table 6: Attacker types suspected by participants.

thieves), the application developer, and remote “hackers”. The number of times each of these entities was mentioned in a response are recorded in Table 6. Thirty-three of these labels come from persons who identified the importance of the key in verifying their partner’s identity but obviously remained skeptical as to the full security of the application.

It is interesting to note that man-in-the-middle attacks were not explicitly mentioned as a possible attack vector in the responses to either of the prompts evaluated here. Impersonation was mentioned frequently in responses to the first prompt, and various tampering by governments and those with physical access to phones and their software were mentioned in responses to the second prompt. The apparent lack of awareness of man-in-the-middle attacks seemed to influence the trust users had in each other’s identity, based on the frequent mentions of things like shared knowledge and videos used when identifying users. Many respondents further demonstrated this unknown attack surface through additional commentary. For example, R24A said he *“just did not consider verifying her identity. Thought [it] would [be] hard to replicate it within this short time.”*

Many users did seem to grasp that there were other attacks possible, but used the term “hacker” as a generic catchall for these. For example, R27B mentioned that no one could read the messages sent between her and her partner *“unless people read over our shoulder or people hack into our Facebook accounts and read them before we delete them.”* Similarly, R36A and R28A stated that the only people who could read the encrypted messages were *“just the two of us unless there were hackers”* and *“not WhatsApp or third parties! But probably people with skills,”* respectively.

In addition to being a catchall, use of the “hacker” response may also be providing insight into belief in a theoretical ceiling of network security by users. Since most users are unfamiliar with the mathematical foundations of cryptography and the details of security protocols, many struggle to adopt secure practices and understand the nature of various threats. On the other hand, users are often aware of their own ignorance in such matters, and these responses might indicate that users account for this in mental models by incorporating a “hacker” entity that is always capable of subverting any piece of the system. In this sense, lack of security knowledge affects both users’ ability to make secure decisions *and* lowers their confidence in security itself.

Some users also expressed some suspicion of the applications themselves for government and/or developer eavesdropping. R24B was suspicious of both: *“Viber (if they want to) & government investigation agencies”*. Others respondents explicitly mentioned “backdoors” built into the applications or general suspicions like R29B: *“I still feel like WhatsApp can read the messages even though they say they can’t.”* Finally, some users were wary of logging, as exemplified by R15A: *“The company I’m sure has records of the texts but [security] depends on if they go through them or not.”*

Overall, the responses indicate that users have a healthy wariness and high-level understanding of impersonation attacks, government and developer backdoors, and physical theft, but that the same cannot be said for man-in-the-middle attacks, both passive and active. It is assumed that some of the mentions of “hackers” refer to this, but these responses were far less specific than for other attacks. In other words, it appears that users’ threat models do not include the ability for attackers to be positioned in between the two endpoints of a conversation. If this was understood, we hypothesize that far less respondents would have relied on physical appearance or shared knowledge as an identity verification mechanism. Since one of the primary goals of the secure exchange of keys is to thwart man-in-the-middle attacks, work may be needed to help users understand this attack vector.

10. CONCLUSION

We used a two-phase study to examine whether users are able to locate and complete the authentication ceremony in secure messaging applications. In the first phase, users were aware only of the need to authenticate and ensure confidentiality, and only two of twelve users were able to locate the authentication ceremony, with an overall success rate of 14%. Participants instead primarily used personal characteristics, such as a person’s voice, face, or shared knowledge. In the second phase, users were instructed about the importance of comparing encryption keys in order to authenticate a partner, leading to an overall success rate of 78%. Users were significantly more successful using Viber. However, the time required to find and use the authentication ceremony was 11 minutes, combined, on average across all applications, which may be so long that it would discourage users from authenticating each other.

Based on our findings, we believe that many users can locate and complete the authentication ceremony in secure messaging applications if they know they are supposed to compare keys. However most people do not understand the threat model, so it is not clear that they will know how important it is to compare keys.

An open question is how secure messaging applications can prompt the correct behavior, even without user understanding. It may be possible to leverage the tendency users have to rely on personal characteristics for authentication. We are exploring the use of social authentication [20] as a way of translating authentication of encryption keys into a method that is more understandable to users.

Another area for future work is improving the authentication ceremony so that it does not take so long to complete. A system like CONIKS [9] may help to automate the process of discovering another person’s key without relying on a single trusted party, while also providing non-equivocation so that key servers cannot deceive users.

11. ACKNOWLEDGMENTS

The authors thank the anonymous reviewers and our shepherd, Lujo Bauer, for their helpful feedback. This material is based upon work supported by the National Science Foundation under Grant No. CNS-1528022.

12. REFERENCES

[1] H. Assal, S. Hurtado, A. Imran, and S. Chiasson. What's the deal with privacy apps?: A comprehensive exploration of user perception and usability. In *International Conference on Mobile and Ubiquitous Multimedia*. ACM, 2015.

[2] E. Atwater, C. Bocovich, U. Hengartner, E. Lank, and I. Goldberg. Leading Johnny to water: Designing for usability and trust. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 69–88, Montreal, Canada, 2015. USENIX Association.

[3] W. Bai, D. Kim, M. Namara, Y. Qian, P. G. Kelley, and M. L. Mazurek. An inconvenient trust: User attitudes toward security and usability tradeoffs for key-directory encryption systems. In *Twelfth Symposium On Usable Privacy and Security (SOUPS 2016)*. USENIX Association, 2016.

[4] A. Bangor, P. Kortum, and J. Miller. Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies (JUS)*, 4(3):114–123, 2009.

[5] A. De Luca, S. Das, M. Ortlieb, I. Ion, and B. Laurie. Expert and non-expert attitudes towards (secure) instant messaging. In *Twelfth Symposium On Usable Privacy and Security (SOUPS 2016)*. USENIX Association, 2016.

[6] S. Dechand, D. Schürmann, T. IBR, K. Busse, Y. Acar, S. Fahl, and M. Smith. An empirical study of textual key-fingerprint representations. In *Twenty-Fifth USENIX Security Symposium (USENIX Security 2016)*. USENIX Association, 2016.

[7] Facebook. facebookmessenger.com. `https://www.messenger.com/`. Accessed: 8 March, 2017.

[8] A. Herzberg and H. Leibowitz. Can Johnny finally encrypt? Evaluating E2E-encryption in popular IM applications. In *Sixth International Workshop on Socio-Technical Aspects in Security and Trust (STAST 2016)*, Los Angeles, California, USA, 2016.

[9] M. S. Melara, A. Blankstein, J. Bonneau, E. W. Felten, and M. J. Freedman. CONIKS: Bringing key transparency to end users. In *Twenty-Fourth USENIX Security Symposium (USENIX Security 2015)*, pages 383–398. USENIX Association, 2015.

[10] S. Milgram and E. Van den Haag. Obedience to authority, 1978.

[11] S. Ruoti, J. Andersen, T. Hendershot, D. Zappala, and K. Seamons. Private Webmail 2.0: Simple and easy-to-use secure email. In *Twenty-Ninth ACM User Interface Software and Technology Symposium (UIST 2016)*, Tokyo, Japan, 2016. ACM.

[12] S. Ruoti, J. Andersen, D. Zappala, and K. Seamons. Why Johnny still, still can't encrypt: Evaluating the usability of a modern PGP client, 2015. arXiv preprint arXiv:1510.08555.

[13] J. Sauro. *A practical guide to the system usability scale: Background, benchmarks & best practices*. Measuring Usability LLC, 2011.

[14] J. Sauro and J. R. Lewis. Average task times in usability tests: what to report? In *Twenty-Eighth ACM Conference on Human Factors in Computing Systems (CHI 2010)*, pages 2347–2350. ACM, 2010.

[15] S. Schröder, M. Huber, D. Wind, and C. Rottermanner. When SIGNAL hits the fan: On the usability and security of state-of-the-art secure mobile messaging. In *First European Workshop on Usable Security (EuroUSEC 2016)*, 2016.

[16] S. Sheng, L. Broderick, C. Koranda, and J. Hyland. Why Johnny still can't encrypt: Evaluating the usability of email encryption software. In *Poster Session at the Symposium On Usable Privacy and Security*, Pitsburg, PA, 2006.

[17] A. Sotirakopoulos, K. Hawkey, and K. Beznosov. "I did it because i trusted you": Challenges with the study environment biasing participant behaviours. In *SOUPS Usable Security Experiment Reports (USER) Workshop*, 2010.

[18] J. Tan, L. Bauer, J. Bonneau, L. F. Cranor, J. Thomas, and B. Ur. Can unicorns help users compare crypto key fingerprints? In *Thirty-Fifth ACM Conference on Human Factors and Computing Systems (CHI 2017)*, pages 3787–3798. ACM, 2017.

[19] N. Unger, S. Dechand, J. Bonneau, S. Fahl, H. Perl, I. Goldberg, and M. Smith. SoK: secure messaging. In *Thirty-Sixth IEEE Symposium on Security and Privacy (SP 2015)*, pages 232–249. IEEE, 2015.

[20] E. Vaziripour, M. O'Neill, J. Wu, S. Heidbrink, K. Seamons, and D. Zappala. Social authentication for end-to-end encryption. In *Who Are You?! Adventures in Authentication (WAY 2016)*. USENIX Association, 2016.

[21] Viber. Viber.com. `https://www.viber.com/en/`. Accessed: 8 March, 2017.

[22] WhatsApp. Whatsapp.com. `https://www.whatsapp.com/`. Accessed: 8 March, 2017.

[23] A. Whitten and J. Tygar. Why Johnny can't encrypt: A usability evaluation of PGP 5.0. In *Eighth USENIX Security Symposium (USENIX Security 1999)*, pages 14–28, Washington, D.C., 1999. USENIX Association.

APPENDIX

A. STATISTICAL TESTS

This section contains the details of the statistical tests we ran.

A.1 Success and Failure Rates

This data measures whether the participants were successful in using the authentication ceremony for each application in the second phase of the study. We want to test whether there are any differences between the applications.

Because the data is dichotomous we used Cochran's Q Test and found that the success rate was statistically different for the applications ($\chi^2(2) = 15.429$, p<.0005).

We then ran McNemar's test to find the significant differences among the pairs of applications. As shown in Table 7, and after applying a manual Bonferroni correction for the three tests (requiring p<0.0167), there is a significant difference between WhatsApp and Viber (p=0.008) as well as between Facebook Messenger and Viber (p<0.0005).

A.2 Task Completion Times

		Fail	Success	N	Exact Sig.
		Viber			
WhatsApp	**Fail**	2	8	48	0.008
	Success	0	38		
		Viber			
Messenger	**Fail**	0	12	42	0.000
	Success	0	30		
		Messenger			
WhatsApp	**Fail**	4	2	42	0.109
	Success	8	28		

Table 7: McNemar's test for success and failure

This data measures the time taken by participants to (a) find the authentication ceremony and (b) complete the authentication ceremony, which was only measured in the second phase of the study. We want to know if there is a significant difference in the time to complete these tasks among the three different applications tested—WhatsApp, Viber, and Facebook Messenger.

We first tested for normality using the Shapiro-Wilk test. As Table 8 shows, the data is not normally distributed for any application ($p<0.05$).

Task	Application	Statistic	df	Sig.
Finding Ceremony	WhatsApp	0.902	38	0.003
	Viber	0.878	46	0.000
	Messenger	0.886	30	0.004
Completing Ceremony	WhatsApp	0.856	38	0.000
	Viber	0.835	46	0.000
	Messenger	0.762	30	0.000

Table 8: Shapiro-Wilk test for task completion times

We next ran the Kruskal-Wallis test, which is a nonparametric test that can determine if there are statistically significant differences between two or more groups. This test rejects the null hypothesis that the distribution of task times is hte same across the applications, for both finding the ceremony ($p=0.031$) and completing the ceremony ($p=0.043$). We next ran pairwise post-hoc tests to determine where the differences occur.

As Table 9 shows, We found a significant difference between WhatsApp and Facebook Messenger for finding the ceremony ($p=0.029$), with Facebook Messenger being faster (mean time, WhatsApp=3.7 minutes, Facebook Messenger=2.5 minutes). We also found a significant difference between Viber and WhatsApp for completing the ceremony ($p=0.021$), with Viber being faster (mean time WhatsApp=8.5 minutes, Viber 6.7 minutes). Note, the significance has been adjusted by the Bonferonni correction for multiple tests.

Task	Comparison	Test Statistic	Std. Error	Std. Test Statistic	Adj. Sig.
Finding Ceremony	Messenger - Viber	14.887	7.616	1.955	0.152
	Viber - WhatsApp	5.492	7.114	0.772	1.000
	Messenger - WhatsApp	20.379	7.926	2.571	**0.030**
Completing Ceremony	Messenger Viber	-12.000	7.702	-1.558	0.358
	Viber - WhatsApp	17.526	7.195	2.436	**0.045**
	Messenger - WhatsApp	5.526	8.016	0.689	1.000

Table 9: Pairwise comparisons from Kruskal-Wallis post-hoc tests for task completion times

A.3 Favorite Rankings

This data measures the system participants selected as their favorite, which was only collected in the second phase of the study. We want to test whether there are any differences between the favorite rankings for each application between the two phases.

We ran a Chi-Square test using the scores for the favorite application. As shown in Table 10, there are no statistically significant differences.

Phase	Favorite WhatsApp	Favorite Viber	Favorite Messenger	Pearson Chi-Square	df	Asym. Sig.
1	9	2	11	2.069	2	0.355
2	15	11	21			

Table 10: Chi-Square test for favorite application ranking

A.4 Trust Scores

We ran a mixed model ANOVA Test because we are interested in seeing the interaction between two independent variables (application and phase). This data is not well suited to a Kruskal-Wallis test because the use of the Likert scale provides too many ties when measuring trust. Mauchly's test of sphericity indicated that the assumption of sphericity was met for the two-way interaction ($\chi^2(2) = 3.385$, p=.184).

We next examined the results for tests of within-subject effects and found that there is a significant interaction between the application and the study phase ($F(2,140)=5.023$, $p=0.008$, partial $\eta^2 = 0.067$).

To determine whether there was a simple main effect for the application, we ran a repeated measures ANOVA on each phase. As shown in Table 11, there was a statistically significant effect of the application on trust for phase 1 ($F(2,46)=4.173$, $p=0.022$, partial $\eta^2 = .154$). Note that due to a violation of the sphericity assumption in phase 2, we use the Greenhouse-Geisser correction.

Phase	Mean WhatsApp	Mean Viber	Mean Messenger	df	F	Sig.	η^2
1	4.13	3.58	3.79	2,46,	4.173	**0.022**	0.154
2	4.10	4.40	4.17	1.69,79.42	1.843	0.171	0.038

Table 11: Repeated measures ANOVA on each phase

By examining the pairwise comparisons, shown in Table 12, we found that the trust score was significantly lower for Viber as compared to WhatsApp in the first phase (M=0.542, SE=0.180, p=0.19). Note, we use the Bonferroni correction for multiple tests.

Comparison	Mean Difference	Std. Error	Adj. Sig	Lower Bound	Upper Bound
WhatsApp-Viber	0.542	0.180	**0.019**	0.076	1.007
WhatsApp-Messenger	0.333	0.155	0.128	-0.068	0.735
Messenger-Viber	0.208	0.225	1.00	-0.373	0.789

Table 12: Pairwise comparisons from one-way ANOVA on each application, phase 1

To determine whether there was a simple main effect for the study phase, we ran a one-way ANOVA on each application to compare the trust between the two phases. As shown in Table 13, there was a statistically significant difference in trust ratings between the two phases for Viber (F(1,70)=14.994, p<0.0005, partial $\eta^2 = .176$). The mean trust for Viber in the first phase was 3.58, and in the second phase it increased to 4.40.

	Mean					
Application	**Phase 1**	**Phase 2**	**df**	**F**	**Sig.**	η^2
WhatsApp	4.13	4.12	1,70	0.007	0.935	0.00
Viber	3.58	4.40	1,70	14.994	**0.00**	0.176
Messenger	3.79	4.17	1,70	2.230	0.140	0.031

Table 13: One-way ANOVA on each application

B. STUDY MATERIALS

This section contains the study materials we used. The interview guide and interview form were used by the study coordinators to ensure that each pair of participants experienced an identical study. The questionnaire was followed by study participants to guide them through the study.

B.1 Interview Guide

Make sure to complete the following steps:

1. When the participants arrive, read them the following:

   Welcome to our secure messaging application study. We are the study coordinators and are here to assist you as needed.

   Before we start the study, we need you to install the following applications: WhatsApp, Facebook Messenger, Viber.

   In this study, the two of you will be in different rooms and will use the applications to communicate with each other. You will each be asked to play the role of another person. I will provide you with information about this person. During the study, please use the provided information and not your own personal information.

   Notice that even you are in separate rooms, you are welcome to ask for meeting, calling or emailing your study partner during the study if you need to complete the study.

   You will be asked to do the task while you are thinking loud and express your feelings or thoughts about each single task that you are doing. During the course of this study we will be recording what is happening in the study room including your any verbal communication with the study coordinators. These recordings will not be seen by anyone beside the researchers and will be destroyed once our research is complete. We will not collect any personally identifying information. Any data, besides the previously mentioned recordings and answers to the study survey, will be deleted automatically upon your completion of the study.

   You will each receive $10 as compensation for your participation in this study. The expected time commitment is approximately 60 minutes. If you have any questions or concerns, feel free to ask us. You can end participation in this survey at any time and we will delete all data collected at your request. A study coordinator will be with you at all times to observe the study and also to answer any questions you may have.

2. Before going to the study rooms, make sure they sign the audio recording consent form.
3. Make sure their phone has enough space for installing the three apps (you can ask them to install the apps before the study starts)
4. Choose one of the available codes for later usage in the study from the following link (a spreadsheet for time slots)
5. Flip a coin and choose one participant to be Person A and one person to be Person B.
6. Take the user with whom you decided to work to the study room. Complete the following setup steps:
   (a) Ask the participant to sit down.
   (b) Start the audio recording using the phones in the lab.
   (c) Read the following instructions to your participant:
   We are going to ask you to do a series of tasks. Once you are done with each step, let the study coordinator know you have finished the task. You will then fill out a questionnaire and go to the next step.
   We need you to think out loud while you are doing the tasks, meaning you are supposed to talk about how you are accomplishing the task and express any feelings you have.
   If you have any questions about the study ask the study coordinator. Remember you are allowed to talk to or meet your friend during the study.
   Please do not forget think loud.
7. On the chromebook, load the survey from Qualtrics
8. Give the code you already selected to the user.
9. Before using each system, the survey will instruct the participant to tell you they are ready to begin the next task.
10. During the course of the task pay attention to what user is doing and fill out one of the attached sheets.
    (a) The user is supposed to think aloud while doing the tasks. If she forgets, gently remind her.
    (b) If the user complains that he is confused, suggest he can consult with his study partner and do not help him to accomplish the task. Try not to instruct the user when they ask questions. Answer them while giving as little information as you can away about the study, but try to remind him that he has a partner who can help him.
    (c) If it takes the pair too long to use one application (10 minutes), then record that as a failure and guide the user to the next task. If you end the task, inform the other study coordinator that you have done so, so that he catches up with you.
11. When the survey is finished, ask the participant about their experience.
    (a) Use the situations you noted while they took the study or interesting things they said on the survey.
    (b) If they had any problems during the study, ask them to use their own words to describe the problem. Ask them how they would like to see it resolved.

12. When the participant is finished, go to meet the other group in your room. Next, ask them the following questions: (If it is applicable)
    (a) You saw QR codes, strings of digits, and maybe NFC communication (touching your phones) as methods for verifying keys. Which one did you prefer and why?
    (b) If you were in a different city or state from your friend, how would you verify your friend's key? Would this be convenient?
    (c) Some of these applications, like Facebook Messenger let you chat both securely and insecurely. The rest of the applications only let you have secure conversations. Which approach do you prefer and why?
13. Thank the participants for their time. Help them fill out the compensation forms. Send them to the CS office to be compensated.
14. Stop the audio recording. Save the record by time.
15. Fill in your name:
16. Return this form.

B.2 Interview Form

Study Coordinator's Name:
Study Number:
System:
WhatsApp, Viber, FaceBook Messenger
Start Time:
End Time:
Key Verification:

- ☐ QR Code
- ☐ Manual verification via phone call
- ☐ Manual verification in person
- ☐ Manual verification other:
- ☐ NFC
- ☐ Verified successfully
- ☐ Notices conversation encrypted

Mistakes Made:

- ☐ The user sends the key or anything related to the key via the application itself
- ☐ The user sends sensitive data (the credit card number) unencrypted or before doing the identity verification
- ☐ Does not use an encrypted conversation
- ☐ Other:

Other:

- ☐ The user calls, texts or meets his study partner Explain:
- ☐ The application crashes and needs to be restarted. Explain:
- ☐ The user expresses any strong feelings toward the task (e.g. how boring or hard or easy it is) Explain:
- ☐ Other Explain:

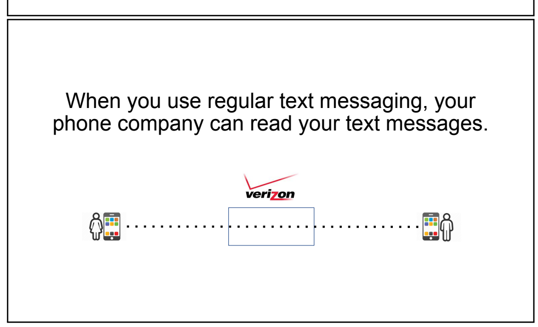
C. STUDY QUESTIONNAIRE

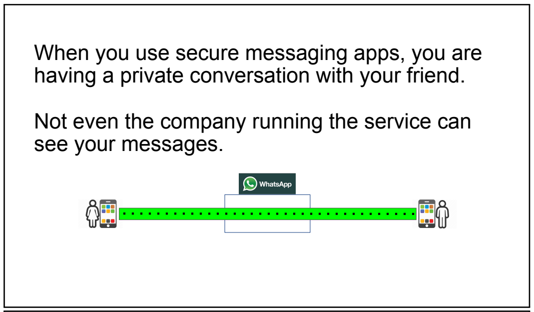
Secure Messaging Application Study

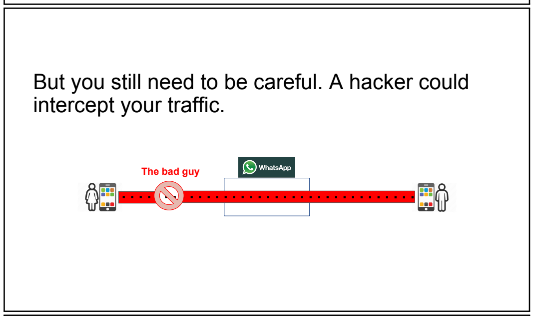
1. Please enter the Type.
    - ○ A
    - ○ B
2. Please enter the code that study coordinator provides for you, here.
3. What is your gender?
    - ○ Male
    - ○ Female
    - ○ I prefer not to answer
4. What is your age?
    - ○ 17 and under
    - ○ 18-24
    - ○ 25-34
    - ○ 35-45
    - ○ 46-64
    - ○ 65 and over
    - ○ I prefer not to answer
5. What is the highest degree or level of school you have completed?
    - ○ None
    - ○ Primary/grade school (2)
    - ○ Some high school, no diploma
    - ○ High school graduate: diploma or equivalent (e.g., GED)
    - ○ Some college, no diploma
    - ○ Associate's or technical degree
    - ○ Bachelor's degree
    - ○ Graduate/professional degree
    - ○ I prefer not to answer
6. What is your occupation or major?
7. Mark any of the following options which apply to you.
    - ☐ Others often ask me for help with the computer.
    - ☐ I often ask others for help with the computer.
    - ☐ I have never designed a website.
    - ☐ I have never installed software.
    - ☐ I have never used SSH.
    - ☐ Computer security is one of my job responsibilities.
    - ☐ I have taken courses related to computer security, electronic engineering, security, or IT.
    - ☐ I often use secure messaging applications such as WhatsApp.
    - ☐ I have never sent an encrypted email.
    - ☐ I am familiar with cryptography.
    - ☐ I understand the difference between secure and non-secure messaging applications.
8. **(Second phase only)** How would you rate your knowledge of computer security?
    - ○ Beginner
    - ○ Intermediate
    - ○ Advanced

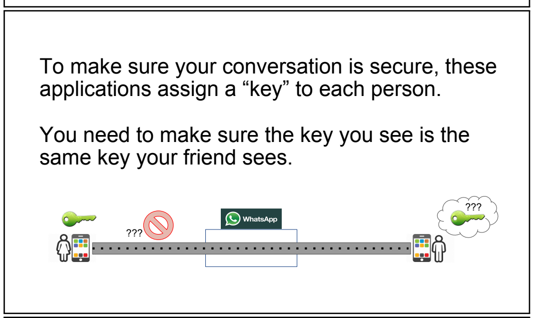
9. **(Second phase only)** If you are reading a website, such as CNN, using HTTP, who can see what you are reading?
   - Nobody, this is a private connection.
   - Your ISP and CNN, but nobody else.
   - Any router in between you and CNN.
   - Your ISP and nobody else.
   - I don't know
10. **(Second phase only)** If you use a regular text messaging application, who can read your text messages?
    - Only the person you send the text message to.
    - The person you send the text message to and the company providing the text messaging service.
    - Anybody who is nearby.
    - Google.
    - I don't know.
11. **(Second phase only)** How can you tell if it is safe to enter your username and password on a website?
    - The website has a good reputation.
    - The website has a privacy statement.
    - There is a lock icon in the URL bar and the URL shows the right host name.
    - The web site is professionally designed.
    - I don't know.
12. **(Second phase only)** What is phishing?
    - Making a fake website that looks legitimate to steal your private information.
    - Hacking someone's computer.
    - Calling someone pretending to be a company to steal their information.
    - Tracking your internet habits to send advertisements.
    - I don't know.
13. **(Second phase only)** What is a public key used for?
    - I do not know what a public key is.
    - To encrypt data for the person who owns the corresponding private key.
    - To setup 2- factor authentication so your password can't be stolen.
    - To identify you to a bank.
    - To protect an application so you know it is safe to use.
14. **(Second phase only)** If you receive a message encrypted with your friend's private key, then you know that
    - Your friend has been hacked.
    - Your friend was the one who sent the message.
    - Everything you send your friend is private.
    - You can't trust what your friend is sending you.
    - I do not know what a private key is.
15. Which of the following applications have you ever used? Select as many options that applies to you.
    - ☐ WhatsApp
    - ☐ ChatSecure
    - ☐ Signal
    - ☐ Telegram
    - ☐ Zendo
    - ☐ SafeSlinger
    - ☐ Allo
    - ☐ FB messenger
    - ☐ iMessage
    - ☐ imo
    - ☐ Skype
    - ☐ Viber
    - ☐ Other
16. What is the main reason why you use these applications (list of applications from previous question) ?
17. Have you ever tried to verify the identity of the person you are communicating with when you are using (list of applications from previous question) ?
    - Yes
    - No
    - Not Sure
18. Have you ever tried to send sensitive information when you use (list of applications from previous question)?
    - Yes
    - No
19. Have you ever had an experience or heard any stories about any secure messaging applications being compromised?
    - Yes
    - No
20. If yes, what story did you hear and what application was it about?
21. **Second Phase Only:**

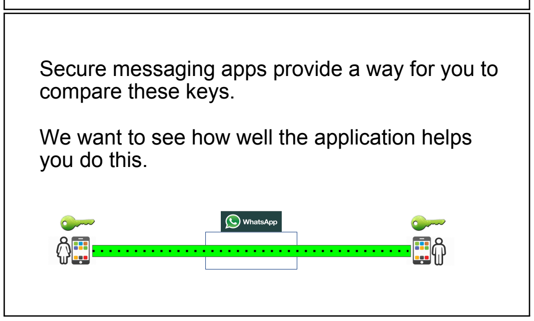
    Read aloud the following instructions:











22. Tell the study coordinator that you are ready for the next task to begin.

    Repeat the following block for each of the three applications

23. You would like to send secure text messages to your friend. For example, you might want to ask for a credit card number you left at home, or talk confidentially about a friend who is depressed.

    In this study we need you to do the following steps:

    For Person A

    You are going to be using (WhatsApp/Viber/Facebook Messenger) for secure texting with your friend. This application is designed to help you have a private conversation with your friend.

    Your task is to make sure that you are really talking to your friend and that nobody else (such as the service provider) can read your text messages. The application should have ways to help you do this.

    We want you to talk and think aloud as you figure this out.

    Once you are sure the conversation is secure, ask the other person to send you your credit card number with the following message.

    “Hello! Can you send me my credit card number that I left on my desk at home?”

    For Person B

    You are going to be using (WhatsApp/Viber/Facebook Messenger) for secure texting with your friend. This application is designed to help you have a private conversation with your friend.

    Your task is to make sure that you are really talking to your friend and that nobody else (such as the service provider) can read your text messages. The application should have ways to help you do this.

    We want you to talk and think aloud as you figure this out.

    Say out loud why you believe you are texting to the right person and why nobody else can read the text messages. Your preference is to figure this out without the other person in the same room, but If you need to visit the other person to do this, you should go ahead and visit them.

    Once you are sure the conversation is secure, he/she will ask you to send his/her credit card number through the application. Use the following number in the study: “132542853779”=

24. You will now be asked several questions concerning your experience with (WhatsApp/Viber/Facebook Messenger).

25. **(Second phase only)** Please answer the following questions about (WhatsApp/Viber/Facebook Messenger). Try to give your immediate reaction to each statement without pausing to think for a long time. Mark the middle column if you don’t have a response to a particular statement.

    - I think that I would like to use this system frequently.
      - Strongly agree
      - Somewhat agree
      - Neither agree nor disagree
      - Somewhat disagree
      - Strongly disagree
    - I found the system unnecessarily complex.
      - Strongly agree
      - Somewhat agree
      - Neither agree nor disagree
      - Somewhat disagree
      - Strongly disagree
    - I thought the system was easy to use.
      - Strongly agree
      - Somewhat agree
      - Neither agree nor disagree
      - Somewhat disagree
      - Strongly disagree
    - I think that I would need the support of a technical person to be able to use this system.

- Strongly agree
    - Somewhat agree
    - Neither agree nor disagree
    - Somewhat disagree
    - Strongly disagree
  - I found the various functions in this system were well integrated.
    - Strongly agree
    - Somewhat agree
    - Neither agree nor disagree
    - Somewhat disagree
    - Strongly disagree
  - I thought there was too much inconsistency in this system.
    - Strongly agree
    - Somewhat agree
    - Neither agree nor disagree
    - Somewhat disagree
    - Strongly disagree
  - I would imagine that most people would learn to use this system very quickly.
    - Strongly agree
    - Somewhat agree
    - Neither agree nor disagree
    - Somewhat disagree
    - Strongly disagree
  - I found the system very cumbersome to use.
    - Strongly agree
    - Somewhat agree
    - Neither agree nor disagree
    - Somewhat disagree
    - Strongly disagree
  - I felt very confident using the system.
    - Strongly agree
    - Somewhat agree
    - Neither agree nor disagree
    - Somewhat disagree
    - Strongly disagree
  - I needed to learn a lot of things before I could get going with this system.
    - Strongly agree
    - Somewhat agree
    - Neither agree nor disagree
    - Somewhat disagree
    - Strongly disagree

26. I trust this application to be secure.
    - Strongly agree
    - Somewhat agree
    - Neither agree nor disagree
    - Somewhat disagree
    - Strongly disagree
27. Have you managed to verify the identity of your friend correctly?
    - No
    - Yes
    - Not sure
28. Please explain why do you think you have (or have not) verified the identity of your friend.
29. Who do you think can read your message except you and your friend?

    End of the repeated block
30. You have finished all the tasks for this study. Please answer the following questions about your experience.
31. Which system was your favorite?
    - WhatsApp
    - Viber
    - FaceBook Messenger
    - I didn't like any of the systems I used
32. Please explain why.
33. Which of the following applications have you ever used for secure communication? Select as many options that applies to you.
    - ☐ WhatsApp
    - ☐ ChatSecure
    - ☐ Signal
    - ☐ Telegram
    - ☐ Zendo
    - ☐ SafeSlinger
    - ☐ Allo
    - ☐ FB messenger
    - ☐ iMessage
    - ☐ Skype
    - ☐ imo
    - ☐ Viber
    - ☐ Other
34. Please answer the following question. Try to give your immediate reaction to each statement without pausing to think for a long time. Mark the middle column if you don't have a response to a particular statement.

    It is important to me to be able to have private conversations with my friends and family using secure applications (like WhatsApp).
    - Strongly disagree
    - Disagree
    - Neither Agree nor Disagree
    - Agree
    - Strongly Agree
35. Did you know about encryption before attending this study?
36. Are you willing to participate in a follow up study? If so, please leave your name and phone number with the study coordinator.

“...better to use a lock screen than to worry about saving a few seconds of time”: Effect of Fear Appeal in the Context of Smartphone Locking Behavior

Yusuf Albayram, Mohammad Maifi Hasan Khan, Theodore Jensen, Nhan Nguyen

Department of Computer Science & Engineering
University of Connecticut, Storrs, CT, USA
yusuf.albayram@uconn.edu, maifi.khan@uconn.edu, theodore.jensen@uconn.edu, nhan.q.nguyen@uconn.edu

ABSTRACT

Using a secure screen lock mechanism is important to prevent unauthorized access and compromise of valuable data stored on smartphones. However, many users still choose not to use any such mechanism and often state inconvenience as the main reason. In this paper, we argue that lack of risk awareness plays an important role behind many users' insecure behavior rather than inconvenience, which can be addressed by communicating risks more effectively. To test this hypothesis, we designed an informational video explaining the risks of unauthorized access to smartphones with no screen lock. We compared a control group ($n = 114$) to a treatment group that viewed the video ($n = 114$) in terms of risk perceptions, concerns, and attitudes towards secure screen lock methods. Subsequently, a follow-up study was conducted to see whether the video was effective in changing participants' behavior or not (i.e., enabling a secure screen lock). We found statistically significant differences between treatment and control group participants in terms of perceived severity, risk awareness, response cost, and privacy and security concerns. Furthermore, the follow-up study revealed that significantly more participants in the treatment group enabled a secure screen lock than those in the control group (48/96 (50%) vs. 21/98 (21%), respectively). Finally, we present our analysis of qualitative data to further explain participants' behavior.

1. INTRODUCTION

Smartphones have become an indispensable part of our daily lives and are increasingly being used for storing and accessing a variety of personal data such as photos, emails, SMS messages, social media posts, and locations. As the sensitivity of the data items stored on smartphones continues to grow, using screen lock features (e.g., PIN, pattern, password, biometric (e.g., fingerprint)) to protect smartphones from unauthorized access is becoming increasingly important [21, 29, 42, 43, 53]. However, prior work has shown that 40%-45% of smartphone users do not use any screen lock mechanism [3, 21] and identified inconvenience (i.e., “It's too much of a hassle”) as the top reason, followed by low perceived data value (i.e.,“No one would care about what's on my phone”) [21, 28].

While these findings may prompt researchers to investigate ways to improve the usability of screen lock mechanisms, which is definitely an important research direction to pursue, we would like to argue that lack of perceived vulnerability and risk awareness can prompt some users to rate the perceived inconvenience relatively higher compared to the perceived benefits of using secure screen lock. Intuitively, if a user does not feel like he/she is at risk of being compromised and does not realize the possible consequences of unauthorized smartphone access, he/she is unlikely to adopt a screen lock mechanism, even if it is convenient and easy to use. On the contrary, if someone realizes the risks of being compromised, he/she is more likely to adopt a secure screen lock mechanism despite minor inconvenience. This argument is in line with findings of prior efforts that have found that when individuals underestimate or are unaware of risks, they often engage in insecure behavior [33, 56, 55]. For example, many users select weak passwords or recycle them from service to service [24] because they think the risk of their accounts being hacked is low [8]. This lack of risk awareness applies to many other security tools and features, including smartphone screen locking [21, 29]. Communicating risks to users is therefore critical for raising their awareness, which in turn can influence their behavior [56, 44].

As such, in this paper, we focus on risk communication using fear appeal and test its effect on users' attitude and behavior. Using guidelines from prior efforts in fear appeal [57], we designed a video focusing on four basic elements, namely, (1) perceived severity (an individual's assessment of the threat), (2) perceived vulnerability (an individual's susceptibility to the threat), (3) self-efficacy (an individual's ability to perform the recommended preventive behavior), and (4) response efficacy (an individual's assessment of the efficacy of the recommended preventive behavior).

Since answering survey questions can raise risk awareness and cause participants to change their locking behavior, to reliably test the effect of the video on users' attitudes and

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

opinions about smartphone screen locks, we ran a controlled study where one group of participants answered the survey questions after watching the video and the other group without watching the video. Participants who completed the main survey were invited to participate in a follow-up survey approximately one week after the completion of the main survey. The aim of the follow-up study was to observe whether or not participants had enabled a secure screen lock method since the first survey. To avoid biasing participants' behavior, we did not inform participants about the possibility of the follow-up in the main study.

In our study, we observed that the video significantly affected perceived severity, vulnerability, and response efficacy. The rating for perceived inconvenience of using the secure screen lock was also significantly lower for the treatment group. Finally, the follow-up study revealed that significantly more participants in the treatment group enabled a secure screen lock than those in the control group (48/96 (50%) vs. 21/98 (21%), respectively).

The rest of the paper is organized as follows. Section 2 presents prior work related to our current study and discusses the basic concepts underlying our analysis. Section 3 explains the study design. Key findings along with detailed analysis are presented in Section 4. Implications of our findings are discussed in Section 5. Limitations of our study along with possible future directions are discussed in Section 6. Finally, Section 7 concludes the paper.

2. RELATED WORK

Use of a secure screen lock is often recommended by security experts [6, 2, 7]. However, many users still choose not to use secure screen locks on their smartphones. Multiple recent efforts in usable security and privacy have tried to understand the reasons behind such insecure user behavior in the context of smartphone security. For example, De Luca et al. [19] and Bhagavatula et al. [13] investigated reasons for (not) using biometric authentication mechanisms on smartphones. Egelman et al. [21] and Harbach et al. [29] investigated why some people choose not to lock their smartphones, finding that inconvenience was a primary reason. Egelman et al. [21] also investigated perception of risk, finding that most participants underestimated the amount of sensitive data that can be mined from their email accounts, and that those who lock their smartphones were more likely to find sensitive data. In the same vein, Harbach et al. [29] found that most users were more concerned about the cost to replace their phone if it was lost than the data on the phone. More recently, Harbach et al. [28] conducted an international survey of 8,286 participants from 8 countries using the Google Consumer Surveys (GCS) platform to investigate whether peoples' smartphone locking behavior differed by country. They found that one third of those surveyed do not use a secure lock screen. Furthermore, the most commonly cited reason was "inconvenience", which was mentioned by 43% (1795) of the participants. In another study, Harbach et al. [27] found that the majority of the participants would not adopt a secure lock screen even if it were more secure. However, most of them indicated that they would be more willing to adopt one if it was quicker than the existing methods. This suggests that "inconvenience" is a major reason for not using any secure lock screen method, and acts as a significant barrier to adoption.

Interestingly, these prior findings are in line with Herley's [30] argument suggesting that users apply a cost-benefit analysis in security-related decisions. Rejection of advice, even that of security experts, occurs when users see the costs of a suggested behavior as too high relative to the benefit. Fagan and Khan [22] recently looked at the risk-cost-benefit perspective of followers and non-followers for different security advices and showed that, in general, each group believes that they are doing the "right" thing. In other words, followers of a certain security advice perceive the benefits of complying as higher than non-followers do. Likewise, followers perceive the risk of not complying as higher than non-followers. This divergence in perception indicates a fundamental difference in the mental models of followers and non-followers, which is unlikely to be resolved by just designing better tools.

One possible way to address this divergence in perceptions is through effective risk communication, which may alter the risk-cost-benefit perspective of non-followers, leading to behavior change. Nudging users in the right direction using approaches such as providing social cues is another alternative. Among a limited number of efforts that investigated ways to motivate users toward adoption of security tools, Das et al. [18] recently examined the influence of social motivations in adoption of recommended security features (login notifications, 2FA, and trusted contacts). They found that 4% of Facebook users adopted one of the recommended security features within a week of seeing an announcement, and 9.9% within five months. Notably, there was no significant difference between the group that viewed the announcement with social cues and the group that viewed it without social cues. More recently, Albayram et al. [9] investigated the effectiveness of informational videos designed to provide an introduction to two-step verification (i.e., 2FA) in persuading users to enable the feature (i.e., behavioral change). They created eight video tutorials based on three themes (i.e., Risk, Self-efficacy, and Contingency) and evaluated the effectiveness of different video content. They also ran a follow-up study to determine whether or not participants who watched the video enabled 2FA for their Gmail account. They found that the Self-efficacy and Risk themes were the most effective in making the videos more interesting, informative, and useful, which were also found to be significantly correlated with participants' decisions to enable 2FA. Moreover, they found that the video including both Risk and Self-efficacy themes had the highest adoption rate, with 39.5% of the participants reporting enabling 2FA for their Gmail account. Van Bruggen et al. [53] used intervention messages with different themes (i.e., morality, deterrence, incentive) to persuade users to adopt screen locks on their smartphones, which is closest in spirit to our work. These were sent as text messages with a 160-character limit. The messages had only a limited effect on behavior (the morality theme was the most effective, with 31% of users changing their locking behavior).

Although these prior studies demonstrated the difficulty in changing users' behavior, none investigated the effect of risk communication in the context of smartphone locking behavior, using fear appeal which emphasizes the potential negative consequences of not following the recommended advice or behavior [57, 36, 15, 49, 51]. According to Witte [57], a fear appeal needs to focus on four basic elements: (1) perceived severity (assessment of threat), (2) perceived vulnerability (susceptibility to threat), (3) self-efficacy (ability to perform the recommended preventive behavior), and (4) response efficacy (assessment of the efficacy of the recommended preventive behavior). Among numerous theories that attempt to explain fear appeal [32, 57, 48], Protection Motivation Theory (PMT) suggests that the motivation to protect depends on threat appraisal (elements (1) and (2) of fear appeal) and coping appraisal (elements (3) and (4)). If threat appraisal results in fear due to an individual assessing a threat as relevant and potentially harmful (i.e., threat appraisal is high), and the individual believes in his or her ability to perform the recommended behavior (i.e., coping appraisal is high), then these two processes elicit protection motivation behaviors [49, 57].

Protection Motivation Theory (PMT) [48, 49] has been applied to study the efficacy of fear appeal in promoting secure behavior such as the use of antivirus software [15] and creation of strong passwords [35] in the past. For example, Boss et al. [15] manipulated the intensity of fear appeal in a longitudinal study, observing that fear appeal had a significant effect on intention to back up data and actual backup frequency. They also found fear appeal to be effective in promoting the use of antivirus software. Jenkins et al. [35] documented the effectiveness of fear appeal in the context of password creation. Specifically, they found that 88% of those who received warning messages when a recycled password was detected created unique passwords, compared to 4.5% of those who did not receive a warning message. Vance et al. [54] examined the impact of fear appeal in motivating users to increase their password strength. During an account-creation process, users were shown one of four messages: one with a fear appeal (i.e., message including susceptibility, severity, self-efficacy and response efficacy), one with interactivity (i.e., showing strength of password as typed-in), both fear appeal and interactivity, and none of these. They found that those shown a fear appeal in interactive form created significantly stronger passwords than those shown only fear appeal or only interactive messages. The effectiveness of fear appeal has been documented in other efforts as well [36, 20].

While a significant volume of prior research has looked at fear appeal in the context of security behavior, to the best of our knowledge, we are the first to investigate the effect of fear appeal on perceived inconvenience and users' attitudes and opinions about smartphone screen locks. Specifically, this study expands upon prior work that has examined why users don't use screen locking mechanisms, and examines the effectiveness of a fear appeal video (designed based on Protection Motivation Theory) in changing users' opinions in terms of perceived severity, risk awareness, response cost, and privacy and security concerns, and persuading them to employ lock screens. The details of our work are presented in the following sections.

3. METHODOLOGY

The goal of this study is to test whether effective risk communication can change users' perceptions of perceived severity, perceived vulnerability, self-efficacy, response cost, and inconvenience, and trigger behavior change in the context of smartphone screen locking. Towards this, we formulated the following hypotheses and designed user studies accordingly.

*H*1: *Participants in the treatment group (i.e., who watched the video) will have higher perceived data value ratings compared to participants in the control group.*

*H*2*a*: *Participants in the treatment group will report being more worried about their smartphones' security and privacy than participants in the control group.*

*H*2*b*: *Participants in the treatment group will report higher level of concerns about their smartphone being used by others than participants in the control group.*

*H*3*a*: *Participants in the treatment group will have higher perceived severity ratings than participants in the control group.*

*H*3*b*: *Participants in the treatment group will have higher risk awareness ratings than participants in the control group.*

*H*4: *Participants in the treatment group will report lower perceived response cost than participants in the control group.*

*H*5: *Participants in the treatment group will have higher ratings for the response efficacy of using a secure screen lock method than those in the control group.*

*H*6: *Participants who watched the video (i.e., treatment group) will be more likely to enable a secure screen lock method compared to those in the control group.*

3.1 Design of the Video

To test the aforementioned hypotheses, we designed a video explaining the risks of not using a secure screen lock. We chose to use video as the mode of communication for conveying fear appeal because it has been shown to be effective in utilizing both visual and auditory information processing channels [52, 17, 40, 31, 46], leading to higher user engagement. In addition, video has been reported to be more effective than text in risk communication [26, 14].

The content of the video was developed based on smartphone security advice from multiple non-profit web sites such as National Cyber Security Alliance (NCSA) [5, 6], EDUCAUSE [2], consumer reports [1] and mobile security companies' blog posts [7, 4]. Over multiple iterations, two researchers narrowed down the risks applicable to most smartphone users (e.g., loss/theft, exposure of sensitive/personal data). The video first explains the possible negative consequences of using a smartphone without a screen lock, and then shows how to set one up on an Android smartphone. Figure 1 shows a frame from the video used in the study. We used screencasts of an Android phone screen to demonstrate some of the possible risks in order to make the video more realistic. Also, as research has identified that aligning security threats with similar risks in the physical world may lead to better understanding by users [25, 10], we used pictures of tools such as locks and shields throughout the video. For self-efficacy, the video demonstrates step-by-step instructions on how to enable a secure lock screen on Android phones[1]. The video transcript was narrated by a native English speaker. No other sound was included in the video. The full video transcript can be found in the Appendix. The full video can be watched on YouTube at `https://youtu.be/J_u6a4ImklM`.

[1]The video presented several available Android screen lock methods so that participants were not manipulated to select a particular one.

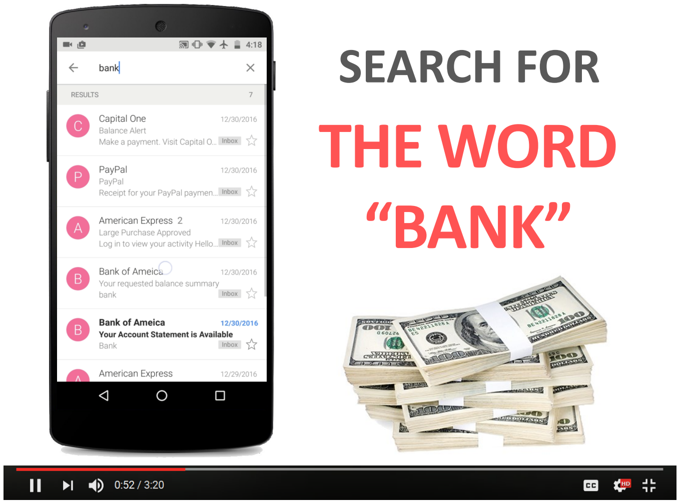


Figure 1: A sample frame from the video is shown. The full video can be watched on YouTube at `https://youtu.be/J_u6a4ImklM`**.**

3.2 Study Design

The study was carried out in two phases: the main survey and the follow-up survey. The flow of the study is shown in Figure 2.

In the main survey, we evaluated the effect of the fear appeal video on users' attitudes and opinions about smartphone screen locks. Half of the total participants (treatment group) were required to watch the video, which was approximately 3 minutes long, while the other half (control group) did not see the video. Participants were assigned randomly to one of these two groups to prevent self-selection bias.

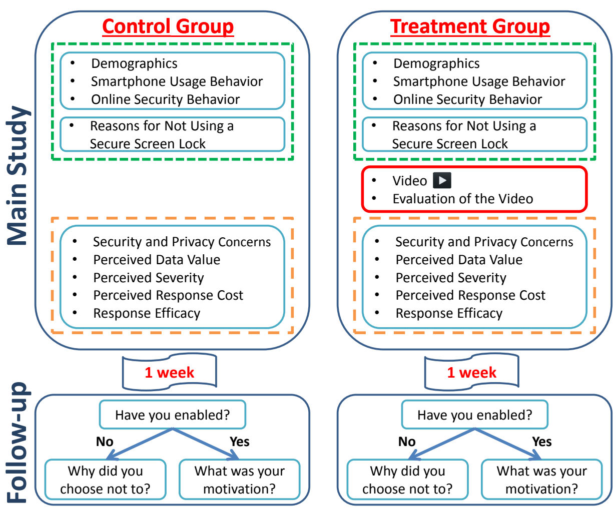


Figure 2: Survey Flow.

The main survey contained three sections. The first section consisted of questions about demographic information, computer and online security behaviors, level of computer proficiency, smartphone usage behavior, and reasons for not using a secure screen lock. Next, participants were assigned to a group. Those assigned to the treatment group were shown the video. We used Qualtrics' timer feature to ensure that participants waited at least the duration of the video (200 seconds) before advancing to the next part of the survey. On average, participants took 201.43 seconds to watch the video (median = 224.3 and SD = 118). After watching the video, participants in the treatment group were asked 6 video-related questions. Four questions asked participants' opinions about the persuasiveness of the video (i.e., its effectiveness in changing behavior and believability). These questions were adopted from prior work by Kazjer et al. [37] and modified accordingly. The fifth and sixth questions were open-ended and asked participants to share what aspects of the video they liked and disliked respectively. The third section consisted of questions about perceived data value, risk perceptions, perceived response cost, and response efficacy. Note that participants in the control group were not asked the video-related questions. Other than these, participants in both groups answered the same set of questions in the same order.

Participants who completed the main survey were then invited to participate in a follow-up survey approximately one week later. The aim of the follow-up study was to measure whether or not participants had enabled a secure screen lock method since the first survey. Participants were sent an email via Mechanical Turk's messaging system that directed them to a new HIT on Mturk and informed them that participation in the follow-up survey was voluntary. If they decided to participate, they were paid $1 regardless of their behavioral outcome. Furthermore, to avoid biasing participants' actions, participants were not informed about the possible follow-up survey during the main survey, eliminating any incentive for lying.

In the follow-up survey, participants were first asked whether or not they enabled a secure screen lock after participating in the main survey. This question served as a branching question. Participants who answered "Yes" were asked "What motivated you to enable the secure screen lock method on your smartphone?" Participants who answered "No" were asked "Why did you choose not to enable a secure screen lock method on your smartphone?"

3.2.1 Recruitment

We recruited participants from Amazon's Mechanical Turk (MTurk) platform. We restricted MTurk workers to those 18 years of age or older, currently living in the United States, having completed at least 1000 HITs and having a 95% or higher approval rating, which is recommended in prior work [45].

We screened potential participants based on the following two criteria.

1. Have a smartphone that runs on Android.
2. Did not use any secure screen locking method (e.g., PIN, Pattern, Password) in the past, and currently do not use one.

As secure screen locking methods may vary depending on mobile operating system, we used the first criterion to choose participants whose smartphones run on Android, making the study tractable. The second criterion allowed us to recruit participants who did not have prior experience with secure screen locking methods. As these individuals are likely to have different mental models compared to those who used a screen lock before, we decided to focus on one group of users rather than mixing them up. To avoid biasing responses to these questions, we did not disclose the eligibility criteria to participants during the prescreening process. Additionally, we asked participants 4 prescreening questions instead of the relevant 2 only, to prevent them from guessing the eligibility criteria. Respondents who did not meet the above criteria were informed that they were not eligible to participate in the study and did not receive any compensation.

In addition to the prescreening questions, we also asked participants whether they use some form of authentication (PIN, Pattern, Password, Biometric (e.g., Fingerprint)) when unlocking the screen on their smartphones, and if they do, which method they use. A total of 3256 respondents completed the prescreening survey and 228 participants were found to be eligible to participate based on our two criteria. Out of 3231 respondents who had a smartphone, a majority of them were Android users (1733, 53%), followed by IOS users (1341, 41%). There were 341 (∼20%) Android users and 168 (∼13%) IOS users who reported not using a secure screen lock on their smartphones. Among the Android users who reported using a secure screen lock, PIN was the most preferred method (∼40%; 550), followed by Pattern (∼33%; 447) and Fingerprint (∼18%; 242). Among the IOS users who reported using a secure screen lock, Fingerprint was the most preferred method (∼44%; 502), followed by PIN (∼39%; 450) and Password (∼11%; 124).

Participants took 15.1 minutes on average (median=12.5 minutes, SD=10.6 minutes) to complete the main survey and 3.6 minutes on average (median=2.7 minutes, SD=4.2 minutes) to complete the follow-up survey. They received $2.5 and $1 respectively for their efforts.

The study was approved by the university IRB board.

3.3 Survey Data Analysis

3.3.1 Statistical Analysis

As our data is ordinal and many of the distributions in our data set are skewed, we use the non-parametric Mann-Whitney test. In the case of comparison tests, we report the effect sizes of significant tests using the $r = Z/\sqrt{N}$ metric [23]. For analyzing the association between two categorical variables, we use Pearson's chi-square test[2] and report ϕ for the effect size [23].

3.3.2 Coding Methodology for Qualitative Data

We coded responses to the open-ended questions using a bottom-up inductive coding approach [41]. Initially, two researchers worked independently and read through all the comments, developing a set of codes for each question. Then, the coders met twice (once for the main survey and once for the follow-up survey) to create the final codebook for each question. A third researcher, who was not involved in the initial coding process, moderated the meetings to help reach agreement on the codebooks. Once the codebooks were finalized, the two coders updated their codebook independently. Inter-rater reliability was calculated for each question using Cohen's *Kappa* which ranged from 0.78 to 0.92, indicating "substantial" or "excellent" agreement between the coders [39]. The details are presented below.

[2] Fisher's Exact test was used when the observed values are less than 5 in the contingency table [23].

4. EVALUATION

4.1 Sample Statistics

We found no significant difference between the control and treatment groups in terms of gender ($\chi^2(2) = 1.01$, $p = .603$), age ($U = 6088$, $p = .410$), level of education ($U = 6144$, $p = .456$), level of knowledge about computers in general ($U = 6201$, $p = .520$), and level of knowledge about computer security in general ($U = 6080$, $p = .362$).

Next, we compared the two groups in terms of their smartphone usage behavior. We found no significant differences regarding the amount of time per day participants spend on their smartphones ($U = 6059$, $p = .370$), number of applications they have installed on their smartphones ($U = 6301$, $p = .692$), and number of applications used on average per day ($U = 6411$, $p = .860$).

We further compared the participants in terms of their security behavior and attitudes. There were no significant differences in participants' concerns about their online accounts being compromised or hacked ($U = 6231$, $p = .568$), whether they worry about their online security ($\chi^2(1) = 2.55$, $p = .146$) and whether they use antivirus software on their smartphones ($\chi^2(2) = 3.44$, $p = .178$).

Based on our analysis, we concluded that the two groups are similar in terms of their demographics, smartphone usage behaviors, security behavior, and attitudes.

4.2 Reasons for Not Using Lock Screen

Participants in both groups were asked why they choose not to use a secure screen locking method. The treatment group answered this question before watching the video, which allowed us to compare the initial reasoning of the two groups. In response to this question, we received a total of 322 comments and organized the comments into 11 codes.

We used the coded responses to this question to identify whether the two groups differ in terms of their reasons for not locking. Performing a chi-square test for each reason (i.e., code), we found no significant differences in terms of reasons for not locking between the control and treatment groups.

Comments in the top four codes mentioned the following: annoyance (e.g., inconvenient, time-consuming) (115 comments; 56 control group, 59 video group), low perceived threat (e.g., my phone is always with me, it is unnecessary) (93 comments; 51 control group, 42 video group), nothing to hide (e.g. no sensitive data on my phone) (45 comments; 24 control group, 21 video group), and setup inconvenience (20 comments; 11 control group, 9 video group). These codes accounted for 84.78% of all the comments. The full coding can be found in Table 6 in the Appendix.

Comments such as *"I have never used it. When I think about it I think about how annoying it is to put in a password in other places like my computer at work and I don't want to deal with that. It takes away from the joy of using my phone. (Video)"* and *"It's a pain in the neck to have to unlock it ev-*

ery time I want to use the phone - very inconvenient (Control)" highlight the response cost many participants associate with using a secure screen lock.

Other comments such as *"I always keep my phone in my pocket or next to me, so there is no chance of me losing my phone or that someone can steal the phone from me. (Control)"* and *"I never leave my cell phone unattended. No one will ever snoop around on my phone because it's always on me except at night when it's charging by my bed. (Video)"* reflect low perceived risk.

Furthermore, comments such as, *"I'm not hiding anything that I don't want people to see. Nothing to hide, no reason to lock the phone (Control)"* and *"I don't feel that I need to use one. I don't have any things on my phone that need to be super secure.(Control)"* reflect low perceived data value, which can also affect perceived severity of being compromised.

Lastly, comments such as *"Just never went to the trouble of figuring it out and implementing it.(Video)"* reflect the initial inconvenience of setting up a screen lock that prevents some participants from implementing one.

Our findings are in line with prior work reporting "inconvenience" (e.g., annoying to use) as the chief reason for not using any screen lock method [21, 29, 28].

4.3 Effect of Fear Appeal on Perceived Data Value

We hypothesize that *participants in the treatment group will have higher perceived data value ratings compared to participants in the control group. ($H1$)*

To test this hypothesis, participants in both groups were asked whether they think the data stored on their smartphones is valuable enough to protect, and how much privacy sensitive data they think their smartphones store.

In the treatment group, 77.2% of the participants responded "Yes" when asked "Do you think that the data stored on your smartphone is valuable enough to protect?", whereas 34.2% responded "Yes" in the control group. Performing a Chi-square test, we found that participants who watched the fear appeal video perceived their smartphone data to be significantly more valuable and worth protecting compared to participants in the control group ($\chi^2(2) = 42.82$, p <.001, $\phi = 0.43$). In addition, participants in the treatment group reported having significantly more privacy sensitive data than those in the control group ($U = 5100$, p =.002, r = .20). Specifically, 49.1% of the participants in the treatment group and 28.1% of the participants in the control group rated either "a moderate amount" or "a great deal" of privacy sensitive information when asked "How much privacy sensitive information do you think your smartphone stores?" on a scale ranging from (1) none at all to (4) a great deal of privacy sensitive information.

This increased awareness was also reflected in participants' comments such as *"...since I watched the video it made me think about security for my phone."* and *"It [Video] pointed out security issues that I was not aware of, such as accessing my email to get bank information and resetting passwords."*

These findings support our hypothesis ($H1$) and underscore the effectiveness of fear appeal in raising data value awareness.

4.4 Effect of Fear Appeal on Security and Privacy Concerns

We hypothesize that *participants in the treatment group will report being more worried about their smartphones' security and privacy than participants in the control group* ($H2a$) and *participants in the treatment group will report higher level of concerns about their smartphone being used by others than participants in the control group.* ($H2b$)

To test these hypotheses, participants were asked to rate how much they worry about their smartphones' security and privacy and how concerned they are about their smartphone being used by others.

We found that participants in the treatment group reported being significantly more worried about their smartphones' security (U = 4413, p <.001, r = .291) and privacy (U = 4361, p <.001, r = .298) compared to participants in the control group. They were also more concerned about their smartphone being used by others (U = 3308.5, p <.001, r = .442). Summaries of responses can be seen in Table 1.

These findings support our hypotheses ($H2a$ and $H2b$) and confirm that participants in the treatment group felt more susceptible to threats (perceived vulnerability), which made them worried about their smartphones' security and privacy.

4.5 Effect of Fear Appeal on Perceived Severity and Risk Awareness

Perceived severity refers to how serious an individual deems a threat and its consequences [48]. If an event is not appraised as likely to occur, or if the negative consequences of an event are considered insignificant, then it is unlikely to cause any change in behavioral intentions or actual behavior [48]. Hence, it is important for an intervention message to increase the rating of perceived risk and severity in order to trigger a change in behavior [48]. As such, we hypothesize that *participants in the treatment group will have higher perceived severity ratings than participants in the control group.* ($H3a$) and *participants in the treatment group will have higher risk awareness ratings than participants in the control group.* ($H3b$).

To test these hypotheses, we asked participants to rate the three statements listed in Table 2. Performing a Mann-Whitney U test, we found that participants who watched the video assessed the seriousness of the loss of data on their smartphone as more disruptive (Mean = 5.21, median = 5) compared to the participants in the control group (Mean = 6.22, median = 7) (U = 5047, p =.003, r = .194) (see Table 2). Moreover, in the control group, 66% indicated being more upset about losing the phone itself and 33% indicated being more upset losing the data on the phone, whereas the treatment group had respective percentages of 51% and 47%, with the difference being significant ($\chi^2(2) = 42.82$, p <.001, $\phi = 0.43$).

Finally, performing a Mann-Whitney U test, we found that participants who watched the video perceived their smartphones as significantly more susceptible to being lost (U = 5186, p =.005, r = .187) and unauthorized access (U = 4126.5, p <.001, r = .332) than control participants (see Ta-

	Control Avg (Med)	Treatment Avg (Med)	Significance Test
How much do you worry about your smartphone's security?	1.94 (2)	2.52 (2)	$U = 4413$, $p < .001$, $r = .291$
How much do you worry about your smartphone's privacy?	2.04 (2)	2.66 (3)	$U = 4361$, $p < .001$, $r = .298$
How concerned are you about your smartphone being used by others?	1.7 (1.5)	2.64 (3)	$U = 3308.5$, $p < .001$, $r = .442$

Table 1: Rating summaries for security and privacy concerns for each group are shown along with U-Tests comparing responses for each statement between the control and treatment groups. Participants answered the first question on a scale ranging from Not at all worried (1) to Extremely worried (5), and the second and third questions on a scale ranging from (1) Not at all concerned to (5) Extremely concerned.

ble 2 for the summaries for ratings). Hence, our hypotheses $H3a$ and $H3b$ were supported.

These changes in perceptions were further reflected in comments from participants in the treatment group as follows.

"I liked that the threats to my information and identity were enumerated including some I had never considered before, like how my contacts could be harmed."

"The way it [Video] highlighted potential harm that could be done by someone using my phone was both frightening and convincing. I never thought about a lot of those things".

These results show the effectiveness of the video in communicating risks to users and changing perceptions critical to behavior change.

4.6 Effect of Fear Appeal on Response Cost

Citing "inconvenience" as the top reason for not using any secure lock screen method suggests that the perceived cost associated with using a screen lock plays a crucial role in users' decisions (see [21, 29, 28] and section 4.2). We argue that perceived inconvenience is relative to perceived risk and severity, and thus can be influenced by effective risk communication. Hence, we hypothesize that *participants in the treatment group will report lower perceived response cost than participants in the control group.* ($H4$).

To test this hypothesis, we asked participants to rate the three statements listed in Table 3 on a 5 point-likert scale ranging from 1 = "Strongly disagree" to 5 = "Strongly agree".

We found that participants in the treatment group rated the statement "If I use a secure screen lock on my smartphone, It will be too much of a hassle for me" significantly lower than the control group ($U = 3980.5$, $p < .001$, $r = .349$) (mean = 3.8 vs. 2.95 and median = 4 vs. 3 for the control and video group respectively). While treatment participants rated the statement "I feel using a secure screen lock on my smartphone is too inconvenient due to entering an unlock code every time I use the phone" significantly lower than the control group ($U = 3625$, $p < .001$, $r = .396$), there was no significant difference in ratings for the statement "I feel using a secure screen lock on my smartphone is too inconvenient because it is hard to remember" ($U = 5555.5$, $p = .052$, $r = .128$). Interestingly, participants did not differ significantly in terms of memorability rating, yet they differed significantly in terms of inconvenience rating.

Participants were then asked to explain the reasons behind their ratings for the statement "I feel using a secure screen lock on my smartphone is too inconvenient due to entering an unlock code every time I use the phone". Participants who rated screen locks as inconvenient gave similar reasons to those previously found. Once again, "annoying to use" was the top code. Of 278 comments that were coded, 112 mentioned this. Notably, 74 comments were from the control group participants (51% of total control group comments) and 38 were from the treatment group (27% of total treatment group comments). There were 72 comments (26% of all the comments) indicating willingness to set up a screen lock. As a testament to the gained risk awareness offered by the video, 57 of these were from the treatment group (41% of the total treatment group comments), while 15 were from the control group participants (10% of total control group comments). The full coding can be found in Table 9 in the Appendix.

The increase in risk awareness in the treatment group was also reflected in the following comments.

"Before taking this survey I would agree with all of those statements. Now that I carefully consider all of the risks and consequences associated with someone gaining access to my phone I'm realizing that it's much better to use a lock screen than to worry about saving a few seconds of time by not locking my phone. I am aware of the bad things that could happen if my phone got lost or got into the wrong hands."

"Before your survey I thought it was too much of a hassle but now I think it's worth it! I really didn't think it through or have enough facts to realize just how much info is stored on my phone or is capable of being easily accessed. It won't take long to set up a secure screen lock AND to use it!"

"I felt that way until I saw the video of potential dangers from not having a secure screen lock. Now I don't think it's inconvenient at all."

"I did think it was too much of a hassle, but after seeing the video I changed my mind. It would be more of a hassle to deal with the consequences of theft of the information."

These comments highlight that the video helped these users to realize the risks and consequences associated with unauthorized access to their smartphones, making them see that the benefits of using a secure screen lock outweigh the cost. This supports our argument that perceived inconvenience can be affected indirectly by raising risk awareness and, thus, $H4$ is supported.

4.7 Effect of Fear Appeal on Response Efficacy

Response efficacy refers to an individual's belief in the benefits of the recommended behavior [48]. Protection motivation theory (PMT) postulates that response efficacy is an important determinant of attitude change [48, 50]. We hypothesize that *participants in the treatment group will have*

	Control Avg (Med)	Treatment Avg (Med)	Significance Test
If your smartphone is lost or stolen, how disruptive the loss of your data on your smartphone will be to your daily life?	5.21 (5)	6.22 (7)	U = 5047, p =.003, r = .194
How likely is it for you to lose your smartphone?	1.69 (2)	2.03 (2)	U = 5186, p =.005, r = .187
How likely is it for someone to attempt to access your smartphone?	1.71 (1)	2.34 (2)	U = 4126.5, p <.001, r = .332

Table 2: Rating summaries for perceived severity and risk awareness for each group are shown along with U-Tests comparing responses for each statement between the control and treatment groups. Participants answered the first question on a scale ranging from little disruption (1) to high disruption (10), and the second and third questions on a scale ranging from (1) Extremely unlikely to (5) Extremely likely.

	Control Avg (Med)	Treatment Avg (Med)	Significance Test
If I use a secure screen lock on my smartphone, It will be too much of a hassle for me	3.8 (4)	2.95 (3)	U = 3980.5, p <.001, r = 0.34
I feel using a secure screen lock on my smartphone is too inconvenient due to entering an unlock code every time I use the phone	4.1 (4)	3.1 (3)	U = 3625, p <.001, r = 0.39
I feel using a secure screen lock on my smartphone is too inconvenient because it is hard to remember	2.8 (3)	2.5 (2)	U = 5555.5, p = .052, r = 0.12

Table 3: Rating summaries for response cost for each group are shown along with U-Tests comparing responses for each statement between the control and treatment groups.

higher ratings for the response efficacy of using a secure screen lock method than those in the control group. ($H5$).

To test this hypothesis, we asked participants to rate the set of statements listed in Table 4 on a 5 point Likert scale ranging from 1 = "Strongly disagree" to 5 = "Strongly agree". As we intended to evaluate actual behavior change in the follow-up study, we decided not to ask about willingness to enable a secure screen lock in the main study, which might nudge participants towards enabling it.

As shown in Table 4, participants in the treatment group agreed significantly more with the statements about the benefits of using a secure screen lock and its efficacy in securing smartphone data. Moreover, they agreed significantly more with the statements about secure screen lock being easy to use (i.e., self-efficacy) and being a good idea. These results suggest that fear appeal positively influenced users' opinions about the effectiveness of secure screen locks. Hence, our hypothesis $H5$ was supported.

These changes in perceptions were also reflected in treatment group comments as follows.

"I liked how the video showed what could happen if someone did take your phone. But then how easy it would be to prevent it if you have a password or other method to keep your information secure."

"I liked the video. It presented valid reasons to lock my smartphone. It's about protecting myself but also my loved ones."

4.8 Effect of Fear Appeal on Behavior Change (Follow-up)

We hypothesize that *participants who watched the video (i.e., treatment group) will be more likely to enable a secure screen lock method compared to those in the control group.* ($H6$).

To test whether the video was effective in changing participants' behavior (i.e., enabling a secure lock screen), we sent an email invitation to participants for a follow-up survey approximately one week after the main study.

Out of 228 participants who participated in the first phase of the study, 194 participants (~85%) responded to our invitation email and completed the follow-up survey (98 from the control group and 96 from the treatment group). In this survey, we asked participants whether they actually enabled a secure screen lock or not. Participants who reported enabling a secure screen lock method were asked which method they enabled and what their motivations were. Participants who reported not enabling any secure screen lock method were asked why they chose not to enable it.

In the follow-up study, 48 participants from the treatment group (50% of those who completed the follow-up survey) reported that they enabled a secure screen lock method on their smartphones, whereas 21 participants from the control group (~21% of those who completed the follow-up survey) reported enabling it. Table 5 shows the percentage of participants who reported enabling a lock screen with respect to their assigned group. We found that there was a significant difference in terms of enabling rate between the control group and the treatment group (50% vs. 21%, $\chi^2(1) = 17.27$, p <.001, $\phi = 0.298$). This result supports hypothesis $H6$.

	Control (n = 98)	**Treatment (n = 96)**
Enabled	21	48
Not Enabled	77	48

Table 5: The percentage of participants who reported enabling a lock screen with respect to their assigned group.

Among the participants who enabled a secure screen lock (69/194), PIN was the most preferred method (55%; 38), followed by Fingerprint (16%; 11), Pattern (14%; 10), Password (55%; 9), and other (1%; 1). Also, 54% reported enabling it on the same day, 17% within 1 day, 12% within 2 days, and 7% within 3 days after watching the video.

	Control Avg (Med)	Treatment Avg (Med)	Significance Test
Using a secure screen lock will be a good idea	3.61 (4)	4.18 (4)	$U = 4130$, $p < .001$, $r = 0.33$
I think a secure screen lock will be easy to use	2.86 (3)	3.55 (4)	$U = 4266$, $p < .001$, $r = 0.30$
I think enabling a secure screen lock will help make my smartphone more secure	3.93 (4)	4.23 (4)	$U = 5235$, $p = .006$, $r = 0.18$
I understand the benefits of using a secure screen lock	4.1 (4)	4.41 (4)	$U = 4834.5$, $p < .001$, $r = 0.24$
I think enabling a secure screen lock will protect my data on my smartphone	3.80 (4)	4.13 (4)	$U = 5084$, $p = .002$, $r = 0.20$

Table 4: Rating summaries for response efficacy for each group along with U-Tests comparing responses for each statement between the control and treatment groups.

All but one participant who enabled a secure screen lock responded "Yes" when asked "Are you still using the secure screen lock method?" in the follow-up survey. This participant (from the treatment group) enabled PIN on the same day he or she watched the video but disabled it after one day. The participant stated the following comment as a reason for disabling: *"It was inconvenient although I'm thinking of putting it back on for safety once again."* All but one participant who enabled a secure screen lock responded "Yes" when asked "Do you plan to continue using the secure screen lock method on your smartphone?".

Participants were also asked to rate the convenience of the screen lock method they enabled on their smartphones. 53% of the participants (37/69) found the screen lock method as either "very convenient" or "convenient", 19% of the participants (13/69) as "neither inconvenient nor convenient" and 27% of the participants (19/69) as either "very inconvenient" or "inconvenient". A majority (13/19) of those who indicated the method as inconvenient had enabled PIN, whereas none of the participants found the Fingerprint method as inconvenient to use.

4.8.1 Reasons for Changing Locking Behavior

In response to the question "What motivated you to enable the secure screen lock method on your smartphone?", we received a total of 78 comments. The top 3 codes mentioned the following: security or privacy concerns (50 comments; 32 video group, 18 control group), the survey (17 comments; 15 video group, 2 control group), and the video (5 comments; 5 video group, 0 control group). These codes accounted for 92% of all the comments. The full list of codes can be found in Table 10 in the Appendix.

Comments such as *"I don't want my personal information be seen by others (Video)"* and *"After doing the last study, I understood the importance of locking my phone to keep my private information safe (Video)"* demonstrate the increase in perceived data value and general risk awareness that can trigger behavior change.

Some participants in the video group specifically mentioned the video as a motivating factor. For example, one participant wrote the following: *"The video in the first survey showed the process of someone stealing the phone, then using the "forgot password" method on a website to have a new password sent directly to the phone. I realized that could easily happen, and it made me concerned enough that I enabled the PIN Protect. (Video)"*. Another participant indicated that *"After watching the video in the survey and learning more about the risks to my security, I changed my initial opinion that it was not really needed or that the risks were not really much of a threat to me. Even though I don't let people use my phone or conduct things like purchases or banking on my phone, I hadn't really thought about using social media and how that can be compromised if someone were to directly steal my phone or if I were to leave it somewhere where it could then be taken. (Video)"*

Comments such as *"the survey motivated me to put a fingerprint lock on my screen and not to be lazy about it (Control)"* and *"I wanted my phone to be less accessible to everyone. (Control)"* show that even answering survey questions might be enough for some participants to consider enabling a screen lock.

4.8.2 Reasons for Not Changing Locking Behavior

In response to the question "Why did you choose not to enable a secure screen lock method on your smartphone?", we received 157 total comments. The top 5 codes mentioned the following: low perceived threat (46 comments; 15 video group, 31 control group), annoying to use (45 comments; 17 video group, 28 control group), lack of motivation (23 comments; 11 video group, 12 control group), nothing to hide (15 comments; 5 video group, 10 control group), and forgot to enable (11 comments; 7 video group, 4 control group). Note the smaller number of comments from the treatment group, possibly indicating the effect of video's risk communication. The full list of codes can be found in Table 11 in the Appendix.

Apart from the reasons found in the main survey, some participants simply forgot to enable secure screen locks. For example, *"Honestly, I completely forgot about it after taking the survey. (Video)"* and *"I just haven't thought about putting a code because it tends to slip my mind. (Control)"*. Some comments (which fell into the "lack of motivation" code) indicated that the participants have not had time to enable a secure screen lock but they intend to enable it (e.g., *"I haven't had time to think about it, but I do intend to do it. (Control)"* and *"I haven't had the time to do it, but I am still planning on setting it up sometime. (Video)"*).

Interestingly, four participants in the control group but none in the video group specifically mentioned lack of knowledge to set it up: *"I still haven't figured out how to do it (Control)."* and *"I have not gotten around to looking into how yet. (Control)"*. This underscores the importance of developing self-efficacy in changing behavior.

4.9 Ratings of the Video

Participants in the treatment group were asked to rate various aspects of the video immediately after watching it.

The results show that participants generally found the video to be believable, persuasive, and effective in changing their opinions about smartphone locking. Specifically, 92% of the participants rated the video as either "very believable" or "somewhat believable", and none of them gave ratings lower than neutral (Mean = 4.62, Median = 5, SD = 0.63, scale ranging from not at all believable (1) to very believable (5)).

In terms of persuasiveness of the video, 89% of the participants found the video to be either "very persuasive" or "somewhat persuasive" (Mean = 4.31, Median = 4, SD = 0.87, scale ranging from not at all persuasive (1) to very persuasive (5)).

Regarding effectiveness of the video in changing participants' opinion about smartphone locking, 73% of the participants indicated that the video was either "very effective" or "somewhat effective" (Mean = 3.87, Median = 4, SD = 1.13, scale ranging from not all effective (1) to very effective (5)).

Finally, 82% of the participants (93) responded "Yes" when asked "Did the video make you more worried about your smartphone's security and privacy?" and 89% (101) responded "Yes" when asked "Did the video make you aware of the sensitive data you store on your smartphone?".

4.9.1 Correlation Between Video Ratings and Actual Behavior Change

To examine whether users' behavior decisions were related to video ratings and attitudes towards using screen lock, we used Spearman's coefficients to analyze the correlations between different aspects of the video and participants' resulting behavior.

We found that participants' ratings about *"Using a secure screen lock will be a good idea"* and whether they actually enabled a secure screen lock or not was significantly correlated ($\rho = 0.305$, $p < 0.001$). Moreover, decisions to enable a secure screen lock were found to be significantly correlated with how persuasive ($\rho = 0.330$, $p < 0.001$) and effective ($\rho = 0.345$, $p < 0.001$) the videos were. The decision to enable was also significantly correlated to ratings for the question "How likely is it that you will experience a situation like the ones presented in the video?" ($\rho = 0.334$, $p < 0.001$).

4.9.2 Likes and Dislikes

In response to the question *"What aspects of the video did you like?"*, we received a total of 137 comments. The comments were organized into 7 codes: explanation of risk (60 comments), information or content (20 comments), simplicity and quality of presentation (17 comments), animation and graphics (16 comments), the demonstration of setting up a screen lock (12 comments), and narration (6 comments). The last code "None" captured 6 comments that were either empty or from a user who did not like any aspect of the video.

Comments such as, *"I liked how it [Video] went over how easily it is for people to manipulate your phone and information"*, *"I didn't like any aspects. It was terrifying to see that this could actually occur starting with your phone being taken and compromised."*, and *"It [Video] showed me how vulnerable I have been by not using a security code."* reflect the effect of the fear appeal included in the video.

We received a total of 117 comments in response to the question *"What aspect of the video did you not like?"*. Out of 117 comments, 75 comments (64%) did not mention any dislikes. The remaining codes included narration (e.g., "The narrator's unenthusiastic voiceover.") (12 comments), upsetting content (e.g., "It made me think of all the bad things that can happen to me if someone gained access to my phone.") (7 comments), irrelevance to the user (e.g. "most of the situations definitely did not apply to me.") (6 comments), and not enough information (5 comments). The full list of codes can be found in Table 7 and Table 8 in the Appendix.

5. DISCUSSION

There are many possible explanations for users' reluctance to use a secure screen lock. Herley [30] suggests that users apply a cost-benefit analysis in security-related decisions and that rejection of advice, even that of security experts, occurs when users see the costs of a suggested behavior as too high relative to the benefit. We see this cost-benefit analysis at play when many of our participants associated inconvenience with using secure screen lock methods (e.g., setting it up, unlocking every time to use the phone, the mental burden of remembering the code). In other words, for many, the perceived response cost of enabling a screen lock outweighs its perceived benefit. This may be because many users see the risks and potential consequences of not using a screen lock as insignificant. Our findings are also in line with prior work that found "inconvenience" (e.g., annoying to use) as the top reason for not using a screen lock [21, 29, 28].

Towards addressing this issue, we found that our fear appeal video was persuasive and effective in changing users' opinions about smartphone locking. A large majority of treatment group participants reported that the video made them more worried about the security and privacy of their smartphone and its data. This was further reflected in their comments reporting the explanation of risks as the most liked aspect of the video. Those who watched the video also reported higher ratings for response efficacy. Moreover, we found that the ratings from the treatment group participants about being worried regarding their smartphones' security ($\rho = 0.296$, $p = 0.003$) and privacy ($\rho = 0.269$, $p = 0.008$) were significantly correlated with enabling a secure screen lock. This is in line with Protection Motivation Theory, which suggests that threat appraisal resulting in fear elicits protection motivation behaviors [48]. In support of this, the follow-up study revealed that 50% of the participants in the treatment group enabled a secure lock screen on their smartphones while only 21% of the control participants enabled. Those who enabled mentioned security and privacy concerns as the major reasons for their adoption decisions. On the other hand, those who did not enable mentioned low perceived threat, inconvenience, and low perceived data value as their primary reasons.

These findings provide evidence that lack of risk awareness plays an important role in users' insecure behavior. As such, risk communication using fear appeals can be an effective way to change users' risk awareness, leading users to behave more securely. The video motivated several users, who initially justified their insecure behavior, to reconsider the risks and potential consequences of not protecting their smartphones. As a result, the perceived security benefits of using a screen lock likely grew to outweigh the costs. Most notably, participants who viewed the video were more likely to enable a secure screen lock method, indicating that communication of risks through fear appeal led to actual behavior change in our case.

These results have a wide variety of applications in the realm of information security. For example, as more organizations employ BYOD (Bring Your Own Device) policies, employees will increasingly use their own devices to store sensitive information. Though organizations may utilize Mobile Device Management (MDM) software solutions that require phone locking, small companies or those with a conservative IT budget may choose not to employ such software solutions in order to reduce costs. In the absence of such MDM solutions, devices with no screen lock or passcode pose significant risks to organizations if they are lost or stolen. Thus, when organizations attempt to promote secure behavior among employees, they can effectively communicate possible risks and consequences of non-adoption of security measures through less costly educational approaches. In general, with increasing use of smartphones worldwide, it is essential that users understand the need for the different security tools available to them. Fear appeals can be an effective way to communicate risks to this growing pool of users and improve their understanding of the privacy and security related implications of their actions.

6. LIMITATIONS AND FUTURE WORK

While this study provides insights regarding the influence of effective risk communication on user behavior and risk perceptions, our work has several limitations as follows.

First, as we compare the video and the lack thereof, the effect observed in our study is a result of all the themes included in the video. Further controlled studies are needed to confirm the effect of different video themes and their interaction (e.g., including and excluding fear appeals and/or self-efficacy).

Second, while we chose to use a video for risk communication and to demonstrate several attack scenarios, some of which are harder to simulate (e.g., lost phone), it cannot be easily personalized targeting individual users. As users often vary widely in their level of motivation, perceptions of threat severity, and self-efficacy [47, 58], the use of this "one-size fits all" approach may limit the effectiveness of persuasive interventions [11, 12, 38]. We strongly believe that the effectiveness of security interventions can be improved by tailoring messages for specific individuals. One possibility is to ask participants to search through their emails looking for sensitive information (as in Egelman's work [21]) that can raise risk awareness and lead to behavior change. As such, future studies should identify characteristics of users that influence the effectiveness of intervention messages promoting the adoption of security tools, and investigate combinations of different approaches for raising risk awareness.

Third, because our study was survey-based, it was not possible to monitor whether or not a participant was actively paying attention to the video. While we required participants to wait at least the duration of the video before advancing in the survey, it is possible that some were multitasking while the video was playing. Thus, we encourage controlled in-lab studies where a participant's engagement could be measured more directly.

Fourth, as users who previously used secure screen locks but stopped using it may have a bias due to negative past experience, we decided to exclude such participants in our study. However, we acknowledge that it is possible that a small number of participants avoided detection in the prescreening phase. As we asked multiple prescreening questions and did not reveal the qualifying criteria, such cases are expected to be rare and are likely to have a minimal effect on our analysis (note that only about 13% of the 1733 Android users screened were found to be eligible to participate in the study).

Fifth, because our study was conducted online, all behavioral change data was self-reported and, thus, unconfirmed. While we tried to eliminate any incentive for lying by providing the same incentive regardless of behavioral outcome and avoid informing participants about the possible follow-up survey, it is possible that some participants misreported their behavior (although we found no such evidence based on participants' responses to the open-ended questions). To address potential bias caused by self-reporting, we encourage field studies where true behavioral changes are measured in situ.

Sixth, while Mechanical Turk users are often more diverse in terms of age, income, and education level, the MTurk population is known to be younger and more tech-savvy than the general population [16]. Furthermore, some participants might be unemployed and prefer to stay at home most of the time. As such, these users may perceive the risk of not having a screen lock as low, as they are likely to feel less vulnerable to unauthorized access and losing their phones. This can limit the generalizability of our results. A larger sample size along with more diverse samples can obtain larger effect sizes, which were "medium" [23] in most of the comparisons in this study.

Finally, this study investigated the effectiveness of a video in communicating risks, which changed the risk perceptions of some users and led them to initial behavioral change (i.e., enabling a screen lock). However, we did not investigate the maintenance of this behavior change. Longitudinal studies are needed examining the long-term behavioral outcomes of fear appeal videos. Furthermore, we tested the effect of the video in the context of screen locking, which may vary for other recommended security advices (e.g., use of 2FA, password manager [34]). Further studies are needed to confirm the strengths and limitations of our approach across different security advices.

7. CONCLUSION

Many smartphone users choose not to use secure screen lock methods, often stating inconvenience as one of the main reasons [21, 29, 28]. As communicating risks to users can be critical for raising risk awareness, which in turn can influence behavior [56, 44], we designed a fear appeal video communicating the risks and potential consequences of not using a screen lock. We evaluated the effectiveness of the video on users' perceptions, concerns, and attitudes towards secure screen lock methods by comparing a treatment group that viewed the video to a control group. Subsequently, a follow-up study determined whether the fear appeal video was effective in changing behavior (i.e., enabling a secure screen lock). We found that the fear appeal video was effective in changing users' opinions in terms of perceived severity, risk awareness, response cost, and privacy and security concerns. In turn, the follow-up study revealed that significantly more participants in the treatment group enabled a secure screen lock than those in the control group (48/96 (50%) vs. 21/98 (21%)). Our findings show that risk communication can effectively change risk perception, which can be a key to promoting secure behavior such as the use of a secure screen lock mechanism on smartphones. We strongly believe that this study provides insights that will enable researchers to design more effective informational videos aimed at increasing security awareness and motivating users to adopt security features and tools.

8. REFERENCES

[1] Consumer Reports : 5 steps to protect your smart phone from theft or loss. `https://goo.gl/fThDO8`. Accessed: 3/3/2017.

[2] EDUCAUSE: Smartphone Security and Privacy: What Should We Teach Our Users? `http://goo.gl/yT5Cz6`. Accessed: 3/3/2017.

[3] Ponemon Research Report 2016: How Much Is the Data on Your Mobile Device Worth? `http://goo.gl/xTv3ea`. Accessed: 3/3/2017.

[4] Smartphone Security Guide: The Easiest Way to Keep Your Phone and Data Safe. `http://goo.gl/2akC0w`. Accessed: 3/3/2017.

[5] Stay Safe Online: MOBILE DEVICES. `http://goo.gl/oAIIZl`. Accessed: 3/3/2017.

[6] Stay Safe Online: SIMPLE SMARTPHONE SAFETY: 6 TIPS! `http://goo.gl/POL6dS`. Accessed: 3/3/2017.

[7] Think like a thief: safeguard your most personal device from loss or theft. `http://goo.gl/CZJ1Dd`. Accessed: 3/3/2017.

[8] A. Adams and M. A. Sasse. Users are not the enemy. *Commun. ACM*, 42(12):40–46, Dec. 1999.

[9] Y. Albayram, M. M. H. Khan, and M. Fagan. A study on designing video tutorials for promoting security features: A case study in the context of two-factor authentication (2fa). *International Journal of Human–Computer Interaction*, pages 1–16, 2017.

[10] F. Asgharpour, D. Liu, and L. J. Camp. Mental models of security risks. In *International Conference on Financial Cryptography and Data Security*, pages 367–377. Springer, 2007.

[11] S. Berkovsky, J. Freyne, and H. Oinas-Kukkonen. Influencing individually: fusing personalization and persuasion. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(2):9, 2012.

[12] S. Berkovsky, M. Kaptein, and M. Zancanaro. Adaptivity and personalization in persuasive technologies. In *Proceedings of the International Workshop on Personalization in Persuasive Technology (PPT'16), Salzburg, Austria*, 2016.

[13] C. Bhagavatula, B. Ur, K. Iacovino, S. M. Kywe, L. F. Cranor, and M. Savvides. Biometric authentication on iphone and android: Usability, perceptions, and influences on adoption. *Proc. USEC*, pages 1–2, 2015.

[14] J. Blythe, J. Camp, and V. Garg. Targeted risk communication for computer security. In *Proceedings of the 16th international conference on Intelligent user interfaces*, pages 295–298. ACM, 2011.

[15] S. R. Boss, D. F. Galletta, P. B. Lowry, G. D. Moody, and P. Polak. What do users have to fear? using fear appeals to engender threats and fear that motivate protective security behaviors. *MIS Quarterly (MISQ)*, 39(4):837–864, 2015.

[16] D. P. Christenson and D. M. Glick. Crowdsourcing panel studies and real-time experiments in mturk. *The Political Methodologist*, 20(2):27–32, 2013.

[17] J. M. Clark and A. Paivio. Dual coding theory and education. *Educational psychology review*, 3(3):149–210, 1991.

[18] S. Das, A. D. Kramer, L. A. Dabbish, and J. I. Hong. Increasing security sensitivity with social proof: A large-scale experimental confirmation. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pages 739–749. ACM, 2014.

[19] A. De Luca, A. Hang, E. von Zezschwitz, and H. Hussmann. I feel like i'm taking selfies all day!: Towards understanding biometric authentication on smartphones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 1411–1414, New York, NY, USA, 2015. ACM.

[20] H. Du, H. Xu, M. B. Rosson, and J. M. Carroll. Effects of fear appeals and point of reference on the persuasiveness of it security communications. In *Intelligence and Security Informatics (ISI), 2013 IEEE International Conference on*, pages 82–84. IEEE, 2013.

[21] S. Egelman, S. Jain, R. S. Portnoff, K. Liao, S. Consolvo, and D. Wagner. Are you ready to lock? In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 750–761, New York, NY, USA, 2014. ACM.

[22] M. Fagan and M. M. H. Khan. Why do they do what they do?: A study of what motivates users to (not) follow computer security advice. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 59–75. USENIX Association, 2016.

[23] A. Field. *Discovering statistics using IBM SPSS statistics*. Sage, 2013.

[24] D. Florencio and C. Herley. A large-scale study of web password habits. In *Proceedings of the 16th international conference on World Wide Web*, pages 657–666. ACM, 2007.

[25] V. Garg and J. Camp. End user perception of online risk under uncertainty. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 3278–3287. IEEE, 2012.

[26] V. Garg, L. J. Camp, K. Connelly, and L. Lorenzen-Huber. Risk communication design: Video vs. text. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 279–298. Springer, 2012.

[27] M. Harbach, A. De Luca, and S. Egelman. The anatomy of smartphone unlocking: A field study of android lock screens. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 4806–4817, New York, NY, USA, 2016. ACM.

[28] M. Harbach, A. De Luca, N. Malkin, and S. 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, pages 4823–4827, New York, NY, USA, 2016. ACM.

[29] M. Harbach, E. von Zezschwitz, A. Fichtner, A. D. Luca, and M. 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 2014)*, pages 213–230, Menlo Park, CA, 2014. USENIX Association.

[30] C. 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. ACM, 2009.

[31] C. Herron, H. York, C. Corrie, and S. P. Cole. A comparison study of the effects of a story-based video instructional package versus a text-based instructional package in the intermediate-level foreign language classroom. *Calico Journal*, pages 281–307, 2006.

[32] C. I. Hovland, I. L. Janis, and H. H. Kelley. Communication and persuasion; psychological studies of opinion change. 1953.

[33] D.-L. Huang, P.-L. P. Rau, and G. Salvendy. Perception of information security. *Behaviour & Information Technology*, 29(3):221–232, 2010.

[34] I. Ion, R. Reeder, and S. Consolvo. "...no one can hack my mind": Comparing expert and non-expert security practices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 327–346, Ottawa, July 2015. USENIX Association.

[35] J. L. Jenkins, M. Grimes, J. G. Proudfoot, and P. B. Lowry. Improving password cybersecurity through inexpensive and minimally invasive means: Detecting and deterring password reuse through keystroke-dynamics monitoring and just-in-time fear appeals. *Information Technology for Development*, 20(2):196–213, 2014.

[36] A. C. Johnston and M. Warkentin. Fear appeals and information security behaviors: an empirical study. *MIS quarterly*, pages 549–566, 2010.

[37] M. Kajzer, J. D'Arcy, C. R. Crowell, A. Striegel, and D. Van Bruggen. An exploratory investigation of message-person congruence in information security awareness campaigns. *Computers & security*, 43:64–76, 2014.

[38] M. Kaptein, P. Markopoulos, B. de Ruyter, and E. Aarts. Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. *International Journal of Human-Computer Studies*, 77:38–51, 2015.

[39] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[40] R. E. Mayer and V. K. Sims. For whom is a picture worth a thousand words? extensions of a dual-coding theory of multimedia learning. *Journal of educational psychology*, 86(3):389, 1994.

[41] M. B. Miles and A. M. Huberman. *Qualitative data analysis: An expanded sourcebook*. Sage, 1994.

[42] I. Muslukhov, Y. Boshmaf, C. Kuo, J. Lester, and K. Beznosov. Understanding users' requirements for data protection in smartphones. In *Data Engineering Workshops (ICDEW), 2012 IEEE 28th International Conference on*, pages 228–235. IEEE, 2012.

[43] I. Muslukhov, Y. Boshmaf, C. Kuo, J. Lester, and K. 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. ACM, 2013.

[44] J. R. Nurse, S. Creese, M. Goldsmith, and K. Lamberts. Trustworthy and effective communication of cybersecurity risks: A review. In *Socio-Technical Aspects in Security and Trust (STAST), 2011 1st Workshop on*, pages 60–68. IEEE, 2011.

[45] E. Peer, J. Vosgerau, and A. Acquisti. Reputation as a sufficient condition for data quality on amazon mechanical turk. *Behavior Research Methods*, 46(4):1023–1031, 2014.

[46] D. Podszebka, C. Conklin, M. Apple, and A. Windus. Comparison of video and text narrative presentations on comprehension and vocabulary acquisition. 1998.

[47] H.-S. Rhee, Y. Ryu, and C.-T. Kim. I am fine but you are not: Optimistic bias and illusion of control on information security. *ICIS 2005 Proceedings*, page 32, 2005.

[48] R. W. Rogers. A protection motivation theory of fear appeals and attitude change. *The Journal of Psychology*, 91(1):93–114, 1975.

[49] R. W. Rogers. Cognitive and physiological processes in fear appeals and attitude change: A revised theory of protection motivation. *Social psychophysiology*, pages 153–176, 1983.

[50] R. W. Rogers and D. L. Thistlethwaite. Effects of fear arousal and reassurance on attitude change. *Journal of Personality and Social Psychology*, 15(3):227, 1970.

[51] D. R. Roskos-Ewoldsen, J. H. Yu, and N. Rhodes. Fear appeal messages affect accessibility of attitudes toward the threat and adaptive behaviors. *Communication Monographs*, 71(1):49–69, 2004.

[52] N. Tempelman-Kluit. Multimedia learning theories and online instruction. *College & Research Libraries*, 67(4):364–369, 2006.

[53] D. Van Bruggen, S. Liu, M. Kajzer, A. Striegel, C. R. Crowell, and J. D'Arcy. Modifying smartphone user locking behavior. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, page 10. ACM, 2013.

[54] A. Vance, D. Eargle, K. Ouimet, and D. Straub. Enhancing password security through interactive fear appeals: A web-based field experiment. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on*, pages 2988–2997. IEEE, 2013.

[55] N. D. Weinstein. Unrealistic optimism about future life events. *Journal of personality and social psychology*, 39(5):806, 1980.

[56] R. West. The psychology of security. *Commun. ACM*, 51(4):34–40, Apr. 2008.

[57] K. Witte. Putting the fear back into fear appeals: The extended parallel process model. *Communications Monographs*, 59(4):329–349, 1992.

[58] I. Woon, G.-W. Tan, and R. Low. A protection motivation theory approach to home wireless security. *ICIS 2005 proceedings*, page 31, 2005.

APPENDIX

A. VIDEO TRANSCRIPT

Hello! This video is designed to explain some of the major risks of not protecting your smartphone and how you can protect yourself from those risks.

As you know, smartphones often store a great deal of information such as emails, SMS messages, bank account information, photos and videos of your loved ones, personal contacts, and locations, some of which can be very sensitive. If a smartphone with no passcode is lost or stolen, it becomes really easy for someone to access the sensitive data stored in it. That's right! Just imagine how easy it would be for someone to pick up your phone and access all your information. For example, if your phone falls into the wrong hands, an attacker can easily search through your email for the word "bank" to figure out whether you do online banking and where you do it. The attacker can then click the "reset my password" link on your bank's website to receive the password reset link on your phone and take control of your bank account. If your email account is linked to many other online accounts, the attacker can use the same technique to take control of your other accounts as well.

Furthermore, the attacker can go through your emails to find other sensitive information, such as your full name, social security number, address, phone number, credit card numbers, date of birth, work related information and so on. If an attacker can obtain these kinds of information, the attacker can piece them together and sell them to identity thieves, or even impersonate you and apply for new credit cards.

If the phone contains compromising pictures of you or your loved ones, the attacker can use them to blackmail you for money or damage your reputation by posting the pictures online or sending them to all of your contacts using the email app on your phone.

Additionally, the attacker can use apps like Facebook messenger, Hangouts or WhatsApp to send messages to your friends or significant others, pretending to be you and asking for money. They can even request them to come to certain places by faking an emergency situation.

These scenarios highlight only a few of the common risks resulting from not locking your smartphone. These can be easily avoided using any of the secure screen lock mechanisms available on your phone such as pin, pattern, password or fingerprint that are strongly recommended by security experts. These are simple security measures to ensure that no stranger can access your smartphone's content without your permission.

Setting up a secure screen lock is easy and usually takes less than a minute. For example, on Android phones, you can go to the "Settings" and from there scroll down until you find the "Security" option. Under the "Security" option, tap the "Screen Lock". You can now select any of the screen lock methods that you want to use.

In this video, we tried to explain the major risks of not protecting your smartphone and how to protect yourself from those risks by enabling secure screen lock on your phone.

We hope that this video helped you to realize the importance of using a secure screen lock mechanism on your smartphone and will encourage you to start using one! Thank you for taking the time to watch this video!

B. CODING TABLES

Reason	Count		
	Control	Video	Total
Annoying to use	56	59	115
Don't see the risk	51	42	93
Nothing to hide	24	21	45
Setup inconvenience	11	9	20
Lack of knowledge for setup	4	8	12
Using other security measures	7	5	12
Mental burden	6	6	12
Haven't considered it	2	7	9
Screen lock is insecure	1	1	2
Others can use my phone if lost	1	1	2

Table 6: Codes' frequency of occurrence in participants' responses to: "Why do you choose not to use a secure screen locking method?" [BEFORE THE VIDEO].

Aspect	Total
Explanation of risk	60
Information/content	20
Simplicity/quality of presentation	17
Animation/illustration/graphics	16
Showing how to set up a screen lock	12
Narration	6
None	6

Table 7: Codes' frequency of occurrence in participants' responses to: "What aspects of the video did you like?".

Aspect	Total
None	75
Narration	12
Upsetting content	7
Not relevant to user	6
Not enough information	5
Too unrealistic	4
Boring	4
Animation/pictures	3
No person	1
Length	1

Table 8: Codes' frequency of occurrence in participants' responses to: "What aspects of the video did you not like?".

Reason	Count		
	Control	Video	Total
Annoying to use	74	38	112
Willing to set it up	15	57	72
Low of perceived threat	20	13	33
Mental burden	14	7	21
Nothing to hide	10	6	16
Not sure/none	5	6	11
Using other security measures	2	2	4
Laziness	2	1	3
Low response efficacy	1	2	3
Haven't thought about it	2	0	2
Share with other people	0	1	1

Table 9: Codes' frequency of occurrence in participants' responses to: "Please explain in a few sentences your choice to the above question (statements about inconvenience of using a secure screen lock).".

Reason	Count		
	Control	Video	Total
Security/privacy concerns	18	32	50
The survey	2	15	17
The video	0	5	5
Other	1	3	4
Bad experience	1	0	1
Social cost	0	1	1

Table 10: Codes' frequency of occurrence in participants' responses to: "What motivated you to enable the secure screen lock method on your smartphone?".

Reason	Count		
	Control	Video	Total
Low perceived threat	31	15	46
Annoying to use	28	17	45
Lack of motivation	12	11	23
Nothing to hide	10	5	15
Forgot	4	7	11
Don't want to change	5	1	6
Lack of knowledge for setup	4	0	4
Mental burden	2	1	3
Using other security measures	2	1	3
Other	0	1	1

Table 11: Codes' frequency of occurrence in participants' responses to: "Why did you choose not to enable a secure screen lock method on your smartphone?".

End User Security & Privacy Concerns with Smart Homes

Eric Zeng, Shrirang Mare, Franziska Roesner
Paul G. Allen School of Computer Science & Engineering, University of Washington
{ericzeng,shri,franzi}@cs.washington.edu

ABSTRACT

The Internet of Things is becoming increasingly widespread in home environments. Consumers are transforming their homes into smart homes, with internet-connected sensors, lights, appliances, and locks, controlled by voice or other user-defined automations. Security experts have identified concerns with IoT and smart homes, including privacy risks as well as vulnerable and unreliable devices. These concerns are supported by recent high profile attacks, such as the Mirai DDoS attacks. However, little work has studied the security and privacy concerns of end users who actually set up and interact with today's smart homes. To bridge this gap, we conduct semi-structured interviews with fifteen people living in smart homes (twelve smart home administrators and three other residents) to learn about how they use their smart homes, and to understand their security and privacy related attitudes, expectations, and actions. Among other findings, we identify gaps in threat models arising from limited technical understanding of smart homes, awareness of some security issues but limited concern, ad hoc mitigation strategies, and a mismatch between the concerns and power of the smart home administrator and other people in the home. From these and other findings, we distill recommendations for smart home technology designers and future research.

1. INTRODUCTION

Anticipated by researchers for some time now, the Internet of Things (IoT) has arrived in the homes of end users. By some estimates, there are already hundreds of millions of connected "smart home" devices in more than 40 million homes in the U.S. alone, and by 2021, that number is expected to double [48, 56]. With the rise of consumer smart home platforms like Samsung SmartThings [55], Apple Homekit [5], and others, as well as connected devices like Amazon Echo [4], Google Home [29], and Philips Hue lightbulbs [45], end users are empowered to set up their own connected, automated, smart homes. These smart homes support desirable features, such as voice-controlled lights and remote-controlled door locks, but they also raise new security and privacy risks.

Indeed, computer security researchers have already identified numerous issues with smart home technology. These issues range from over-privileged applications running on smart home platforms [26] to viral attacks that can spread between infected lightbulbs [50]. The recent Mirai malware — which compromised connected devices and conscripted them into a botnet, disrupting the internet for millions of people [43] — shows that these risks are already leading to concrete attacks. We discuss additional examples in Section 2.

However, despite an increased focus on smart home security, and the reality of the emerging risks, there has been little study of the security and privacy concerns of *end users* who set up and use these smart home platforms and devices. Without an understanding of the concerns, needs, and use cases of these end users, researchers and smart home platform designers can neither prioritize which problems to focus on, nor develop effective solutions.

We aim to bridge this gap in this work, asking questions such as: how and why do people use their smart homes? What sorts of mental models have users developed for smart homes? What are their security and privacy concerns (or lack thereof), and how do these compare to the risks identified by security researchers? What sorts of issues play out in homes with more than one user? What security or privacy mitigation strategies do end users already use, and where are additional technical solutions or other design efforts needed?

We explore these questions through in-depth, semi-structured interviews with fifteen participants. All participants live in smart homes: twelve administer their smart homes, and three live in a smart home administered by someone else.

We find that our interview participants have an assortment of (generally sparse) threat models, and that the sophistication of their threat models often depends on their technical knowledge of smart homes. And while participants identified security and privacy issues such as data collection, surveillance, or hacking, most were not concerned about these issues on a day-to-day basis. We also identify tensions that can arise in smart homes with multiple users, which in the extreme could lead to potentially dangerous situations if the administrator of the smart home uses the technology to spy on or deny access to other users.

From our findings, we distill lessons and recommendations for future smart home platforms and devices. For example, we recommend further studying and designing consciously for multi-user interactions in smart homes, and we recommend improving user awareness and control through careful UI/UX design, including the inclusion of physical controls on devices. Ultimately, better understanding end users will help us identify gaps between current system designs and users' security needs and expectations, as well as tensions between users' functionality and security needs, and will help focus the efforts of the research community and industry.

In summary, our contributions include:

1. We conduct in-depth, semi-structured interviews with fifteen smart home users, studying how and why they use smart home technologies, their mental models, their security and privacy concerns (or lack thereof), and the mitigation strategies they employ.
2. Among our findings, we learn that participants' threat models are sparse and depend on the sophistication of their technical mental models, that many current smart

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

home users are aware of potential security and privacy issues but not generally concerned, and that tensions may arise between multiple residents in a smart home.

3. From these findings, we distill recommendations for the designs of future smart home platforms and devices, as well as identify opportunities for future work.

2. BACKGROUND AND RELATED WORK

The Internet of Things (IoT) is a broad term for internet-connected devices, which has come to encompass everything from connected cars, wearables, and connected industrial/manufacturing equipment. Our focus is on *smart home* technology, which we consider to include internet-connected appliances, lighting, sensors, door locks, and other objects designed for the home environment. This technology enables applications like security systems and remote monitoring, lighting and climate control that adapts to a user's presence and habits, and voice controls for lighting and appliances.

Current Smart Home Technology Landscape. In recent years, we have seen a rapid increase in the number and type of consumer-oriented internet-connected devices for automating home environments. While home automation technology has existed for decades, smart home devices are now internet connected, interoperable between different vendors, and controllable via smartphone.

Standalone smart devices include thermostats (e.g., Nest), lights (e.g., Philips Hue), motion detectors, door/window sensors, air quality sensors, power outlets, and door locks. Some of these devices connect to the internet through existing Wi-Fi networks, while others use low energy protocols like Zigbee and Z-Wave, and communicate to the internet through a bridge. Smart devices allow users to *automate* their home, e.g., automatically adjusting the thermostat, or turning on or off lights based on motion sensor readings.

Two types of smart home platforms have emerged: *hubs* and *cloud-based integrations*. Hubs — such Samsung SmartThings [55], Wink [2], and Vera [1] — are central hardware devices that other smart home devices communicate with, and can act as a Z-Wave or Zigbee bridge. Via the hub's companion app or website, users can program *automations*. Some hubs, like Samsung SmartThings, support *third-party apps*, which are prepackaged, complex automations written by other developers. Similar to hubs, emerging intelligent personal assistants, like the Google Home and Amazon Echo, can be integrated with many existing smart home devices, allowing users to control their smart home using their voice.

On the other hand, cloud-based integrations rely on the fact that for many stand-alone devices, commands from a user's phone to the device transits the cloud. These cloud services often expose APIs for controlling devices over HTTP. Middleman cloud services like IFTTT (If This Then That) and Stringify can use these APIs to connect stand-alone devices together, and to run automations.

Smart Home Security and Privacy Concerns. Security experts have raised concerns about the security and privacy risks with internet-connected devices in homes [6, 30, 53]. Concerns include privacy risks due to pairing and discovery protocols that leak information about devices in the home [62], insecure communication leaking sensitive information about the home and the residents [17], and vulnerabilities in the devices that can allow an attacker to remotely spy on residents or disrupt their lives [21, 22, 44]. Technological solutions when not implemented correctly may amplify social issues [58]. Shared in-home devices presents new access control challenges [59], which, if not addressed carefully, may amplify interpersonal issues among residents.

Researchers have begun analyzing smart home platforms and devices (e.g., [24, 26, 44]). Findings include over-privileged applications on smart home platforms and vulnerable devices like locks [32] and lightbulbs [42, 50]. Attacks have also occurred in the wild: the massive Mirai DDoS botnet attack disrupted the internet for millions of users [43], a glitch in the Nest thermostat left users in the cold [8], a baby monitor was hacked and a vulnerability in Foscam cameras left thousands of users vulnerable to similar attack [31], and recent reports suggest that internet-connected smart TVs can be used to record conversations [52]. Furthermore, a recent report indicates that IoT malware and ransomware attacks are on the rise [38]. In response to these concerns, researchers have begun to develop designs for more secure smart home platforms (e.g., [27, 54, 63]).

End-User Studies. Prior research on end users of smart homes has generally not focused on security and privacy issues but rather on usability issues, such as installation, motivations and use cases, and the interfaces for control and automation. [10, 20] Research in this area has identified tensions that arise due to differences between members of the household. Brush et al. and Mennicken et al. found that there is often one user who is most enthusiastic and others who interact with the smart home more passively [9, 41]. Ur et al. studied differences in privacy attitudes between teens and parents regarding home-entryway surveillance [60]. Mennicken et al. implemented a calendar based interface for smart home configuration to make it more accessible to passive users [40]. Our work surfaces a similar dynamic between primary and incidental smart home users.

Some prior work has also investigated security and privacy concerns of end users. Brush et al. [9] visited 14 smart homes to study adoption issues, and among their findings, found concerns about security-critical devices like smart door locks and cameras. Worthy et al. [61] asked five subjects to keep an ambiguous IoT device in their homes for a week, finding trust as a critical factor in IoT technology acceptance. Choe et al. [12] asked 22 participants to take devices home for four weeks and studied their perceived benefits and concerns, finding more concern than we do in our study.

Our research contrasts with prior work in three ways: first, we interview participants who have been living in a smart home for months, past the novelty phase and into day-to-day use. Second, we focus primarily on security and privacy, rather than general usability issues. And lastly, we contribute an updated understanding of usability, security, and privacy issues for the current generation of smart home devices, such as Samsung SmartThings, Amazon Echo, and Philips Hue.

Further afield, others have studied security and privacy concerns of end users for related technologies, including parent-child interactions with connected toys [39], security and privacy issues with household robots [11, 22], access control challenges in the home [37], and privacy issues with using smart home technology for assisting senior citizens [16, 57].

3. RESEARCH QUESTIONS

To inform the design of more secure smart homes in the future, we set out to investigate the following research questions.

General Smart Home Use. We ask: What are the common use cases for smart homes today? While the types of home IoT devices have proliferated in recent years, ranging from smart egg trays to smart dolls, learning which types of devices, platforms, and automations are typically present in smart homes will help us understand which security and privacy issues are most salient in this space, and which functionality or other factors are critical to users.

Smart Home Technology Mental Models. We ask: What mental models do users have of their smart home? For example, do their mental models include communication between devices in the home, and/or communication beyond the home (i.e., in the cloud)? Prior work has found that incomplete mental models about a technology leads to incomplete threat models and limited adoption or use of security tools (e.g., email encryption [49], internet privacy [34]).

Smart Home Threat Models. We aim to learn about the specific threat models and security concerns — or lack thereof — of smart home end users. Experts have developed extensive threat models for IoT and smart homes, informed by a technical understanding of the potential vulnerabilities. End users may develop different threat models. We investigate the potential gap between a security expert's threat model and what users are concerned about. What risks are users unaware of or unconcerned about, and are experts considering all of the issues that matter to end users?

Mitigation Strategies. As part of studying end user threat models, we also investigate any mitigation strategies they use when they do have security or privacy concerns. For example, do users change their in-home behaviors around devices that record audio or video? If they employ technical mitigation strategies, are these strategies actually effective?

Multi-User Interactions. What unique security or privacy issues arise in smart homes due to their shared nature? Today, people increasingly use personal computing devices that are not shared with others, like laptops or smartphones [35]. However, smart home technologies are located in common spaces and are critical to basic functions of the home, such as lighting or physical access, thereby affecting all residents. We explore whether incidental users of smart homes, who were not primarily involved in the system's configuration, hold different security and privacy concerns than the primary user, or view the primary user as a potential adversary.

Other Constraints and Requirements. In addition to security and privacy factors, we anticipate that participants will make choices about whether and how to set up their smart homes based on other factors, including convenience, functionality, usability, reliability, and latency. These constraints and requirements may affect what security and privacy solutions are acceptable for end users.

Recommendations for Researchers and Smart Home Designers. Through this investigation, we aim to develop recommendations for smart home designers and for researchers. Specifically: Where should the computer security community focus its efforts? Given the range of potential issues to address, what type of work should be prioritized, and how? For example, should we prioritize better protecting users from malicious or misbehaving third party automations? How can we design devices to promote better mental models and security behaviors? We return to these questions in Section 6.

4. METHODOLOGY

In this section, we describe our study methods and materials.

4.1 Pilot Interviews

Before designing our interview questions, we conducted an exploratory interview with a colleague who set up and lives in a smart home. After designing the initial interview questions, we conducted four additional pilot interviews with smart home residents, and made modifications to the questions to improve their clarity, and to better answer our research questions. We do not include exploratory or pilot interview data in our general results, though we present one particularly relevant anecdote from one of these interviews.

4.2 Recruitment and Screening

We recruited participants by advertising on relevant mailing lists, on smart home related Reddit communities, and via the researchers' social media accounts (Twitter, Facebook).

Potential participants were asked to fill out a screening survey, selecting which, if any, smart home platforms or devices they own, how long they have been using their smart home, whether they set it up themselves, how many other people live in the home, as well as demographic information (age, gender, profession). Participants were also asked to provide their name and email address if they were willing to participate in a phone or Skype interview. We used the screening responses to select participants with at least one smart home platform and covering a range of technical skill levels (inferred from profession); we also explicitly recruited and selected participants who used but did not set up or manage their own smart home.

Participants who completed the phone or Skype interview were compensated with a $10 Amazon gift card; participants who filled out only the survey did not receive compensation.

4.3 Interview Procedure

Participants who were selected for the full interview were then contacted by the researchers to schedule a phone or Skype call. Interviews were conducted by two researchers: one leading the interview and another taking notes and recording the session. We asked participants about:

General Questions: We asked participants to describe the smart home devices they own, how they use them, what apps or automations they have installed, and whether they access these devices remotely or only while physically in the home.

Mental Models: To elicit participants' mental models and degree of technical understanding of their smart home, we asked them to explain how their smart home works, verbally and through a drawing exercise. Drawings have been found to be an effective method for externalizing mental models in conjunction with verbal reports [33], and has been used in several studies of the relationship between mental models and security [34, 47, 49].

We allowed participants to either create a diagram electronically using Google Drawings, or to draw on paper and send us a photograph. We show examples in Section 5.

Security Concerns: In order to avoid prompting participants to merely agree with the interviewer that security and privacy concerns might arise with smart homes (i.e., avoid participant response bias [7, 19]), we began by asking more general questions that could elicit security or privacy concerns but did not explicitly mention them. We asked whether they had hesitations about getting any of their smart home devices, whether there were any devices they thought about getting but ultimately decided against, or whether there were any devices they used but later deactivated.

For participants who did not organically bring up security or privacy concerns, we then prompted specifically about security and privacy (making it clear that a lack of such concern was a valid response, again to avoid participant response bias). We also asked if they had heard about security and privacy concerns with smart homes in the news, and whether they shared those concerns or felt they were overblown; and we asked participants to compare their concern about smart homes to their concern about phones and laptops.

Mitigation Strategies: We asked participants whether their security and privacy concerns (if any) had caused any changes in behavior (e.g., acting differently around smart home devices or changing device settings).

Multi-User Scenarios: We asked participants how many people live in their home, who has what types of access to the smart home, whether they have had disagreements with others about the smart home, and whether house guests have interacted with the smart home.

Technical Skill: We asked participants to self-report, on a scale of 1 (novice) to 5 (expert), familiarity with technology in general, smart home technology, and computer security.

Wrap-Up: Finally, we asked participants if there were any questions they expected us to ask, and gave them a chance to tell us anything else about their smart home.

As an in-depth, qualitative interview, we tailored our questions to the context of individual conversations. Thus, although all participants were asked the above questions, we also asked relevant follow-up questions where appropriate. A copy of the interview protocol is provided in Appendix A.

4.4 Data Analysis

We used a bottom up qualitative method to analyze the data. Three researchers independently read notes from the interviews and listened to recordings, and generated list of themes. Then, the researchers met in person to consolidate the most salient themes into a shared codebook, which consisted 16 structural codes (based on our research questions), further divided into 116 subcodes. The structural codes were broad categories, such as "Mitigation Strategies", and the subcodes enumerated specific instances mentioned by participants, e.g. "Network segmentation". Then, each interview was independently coded by two of the three researchers. One researcher was the primary coder, and participated in coding each interview. After all interviews were coded, the researchers resolved disagreements resulting from human error or misunderstanding of the codes, where possible. Cohen's kappa, a measure of inter-coder agreement, was 0.96. (Fleiss rates kappa values over 0.75 as excellent agreement [28].) Since there are some remaining disagreements, in Section 5, we report numerical values based on the primary coder.

4.5 Ethics

This study was reviewed by our institution's IRB, and was considered exempt. We did not ask participants to reveal sensitive information like account names or home addresses. All participants provided informed consent to participate in the study and be audio-recorded. We stored all interview recordings in password-protected form and removed any identifying information from notes and transcripts.

5. RESULTS

We now turn to a discussion of our results, organized according to the research questions presented in Section 3.

5.1 Participants

Thirty-three participants completed the pre-screening survey, and we conducted interviews with 15 of them, selecting people with smart home platforms and devices, and covering a range of technical skills and other factors. Interviews were conducted in Feb. 2017 and lasted on average 38 minutes.

Of the 15 participants (summarized in Table 1), four were women, eight did not mention having a background in IT or computer science, and two were aged 55 years or older. Participants had smart homes for at least two weeks and up to eight years. Table 1 presents self-reported familiarities with technology, security, and smart homes. However, in some cases, these self-estimates seemed miscalibrated. For example, one participant reported only a "3" in technology familiarity, but was able to describe a cloud-based client-server architecture for smart homes, while others who reported high familiarity with security did not articulate specific concerns even when directly asked. Nevertheless, we include these values as rough indicators of confidence in their abilities.

5.2 General Smart Home Use

We begin by describing the smart home devices participants own, how they use these devices, and how they orchestrate automations between the devices. These details will provide context for subsequent results, and they highlight use cases that computer security solutions must take into account.

Devices. Participants reported having a large variety of internet-connected devices, from many different manufacturers. We summarize these devices in Table 2. Most common are smart lights, thermostats, cameras, and switches. Participants using their smart homes as security systems typically had sensors on doors and windows, as well as motion sensors.

Nine participants mentioned having a hardware hub, and a few others mentioned using apps for centralized control, like Apple Homekit. Intelligent personal assistants, such as the Amazon Echo or Google Home, are also very common (13).

Some of the more uncommon smart devices were custom-made by the participants. For example, P8 was able to automate the lights and jets on their swimming pool by integrating its control systems with a Raspberry Pi, and implemented custom software to decode the data stream and integrate it with their SmartThings hub.

Use Cases. We identified four common smart home use cases: increasing physical safety (including security systems, door locks, and smoke detectors; 9 participants), home automation (automatically adjusting lighting, temperature, or other devices; 13 participants), remote control, and in-home sensing. Many participants mentioned multiple use cases.

ID	Gender	Age	Profession	Primary User?	CS/IT Background?	Self-Reported Familiarity with... Technology	Computer Security	Smart Homes
P1	Male	35-44	IT Security	Yes	Yes	5	5	3-4
P2	Male	35-44	Marketing	Yes	Yes	4	4	3
P3	Female	55+	Biologist	Yes	No	2-5	3.5	3.5
P4	Male	25-34	Healthcare IT	Yes	Yes	4	4	3
P5	Male	25-34	IT Technician	Yes	Yes	5	4	4
P6	Male	25-34	Engineering PM	Yes	Yes	5	4	5
P7	Male	25-34	Fundraiser in higher ed	Yes	No	4	3	5
P8	Male	45-54	Software Engineer	Yes	Yes	5	4	4
P9	Male	25-34	Finance	Yes	No	4	4	3
P10	Male	55+	Chief Financial Officer	Yes	No	4	4	4-5
P11	Male	55+	Professor	Yes	No	3	3	3
P12	Male	18-24	Retail supervisor	Yes	No	5	4	4
P13	Female	18-24	Student	No	Yes	3	2	2-3
P14	Female	25-34	Academic Admin	No	No	2	1	2
P15	Female	18-24	Student	No	No	3-4	3	3

Table 1: Summary of participants. Familiarity was self-reported on a scale of 1 (low) to 5 (high).

Type of device	Count	Examples
Lights	15	Philips Hue, Belkin Wemo Link, Osram Lightify, HomeBrite, LIFX
Intelligent Personal Assistant	13	Amazon Echo, Google Home
Thermostat	12	Nest Thermostat, Emerson Sensi, Ecobee Thermostat
Camera	11	Nest Cam, Withings Home, Foscam, Ubiquiti Aircam
Power outlets and switches	10	Belkin Wemo, Lutron Caseta
Motion Sensor	10	—
Hub	9	Samsung SmartThings, openHAB, Vera, Abode
Door Lock	7	Kwikset Smart Lock
Smoke detector	4	Nest Protect
Leak detector	2	—

Table 2: Devices owned by participants. Only devices owned by more than one participant are listed.

Though remote device use opens the door for security and privacy risks, we find that it is a critical feature for many users: nine participants remotely controlled devices, like lights or thermostats, while eleven used devices to remotely sense within the home, including monitoring things like camera feeds, air quality, and status of devices.

Modality. Participants interact with smart home devices in several different ways, often in combination. 14 of 15 participants use a smartphone app to control or program their devices. 13 participants use contextual triggers, i.e., behavior that executes based on the context, like the time of day or whether the user is home. 12 participants use an Amazon Echo or Google Home to control their devices via voice. 8 participants mentioned using motion sensors. Some remarked that using mobile apps was tedious, and preferred to use voice controls or automations exclusively.

Automations. We define *automations* to be programs that cause devices to do something on their own, or programs that connect two different types of devices so that one can trigger the other (e.g., enabling voice-controlled lights by integrating them with the Amazon Echo). Furthermore, we distinguish between three types of automations: end user programming, custom scripting, and third party apps.

Most standalone devices and hubs feature an end user programming interface, which allows users to program automations for their home on a graphical interface, usually in a mobile app. For example, the SmartThings mobile app allows users to program "routines" for devices like lights: users can trigger lights to turn on and off based on activity from motion sensors, door sensors, time of day, or whether their phone is present in the house. We found end user programming to be the most common method for automations; 11 of 15 participants used this type of interface.

Four more technically skilled users automated their homes by writing scripts for Raspberry-Pi based controllers, like openHAB or HomeAssistant. Three others used custom scripts written by *others*: P7 and P10 downloaded scripts from smart home forums, and P14's openHAB was programmed by her husband. P10 was actually able to request others to write Vera automations for him, and when we asked him about it (incredulously), he said, "Yeah, isn't that great? I've done it 3 or 4 times." Though code taken directly from others may pose security risks, he was not concerned about this risk, as we discuss further below.

Devices can also be automated by third-party tools, such as apps on appified platforms like SmartThings, or cloud-based tools like IFTTT. These methods are used (1) to provide complex automations not possible through end user programming, like adjusting the thermostat based on outdoor temperature, and (2) to integrate devices that are not built-in to a platform, like connecting an Echo to SmartThings.

We found that third-party automations were less common than custom programming solutions. Four participants mentioned cloud services like IFTTT and five mentioned using app-based automations. Both were mostly used when hubs did not provide sufficient integration or functionality with certain device families. Two non-integration automations mentioned were a disco light app for Philips Hue, and a door lock code management app for SmartThings. As we discuss further in Section 6, this finding suggests that research efforts focusing on the security of smart home applications (e.g., [26]) may be considering only a narrow use case.

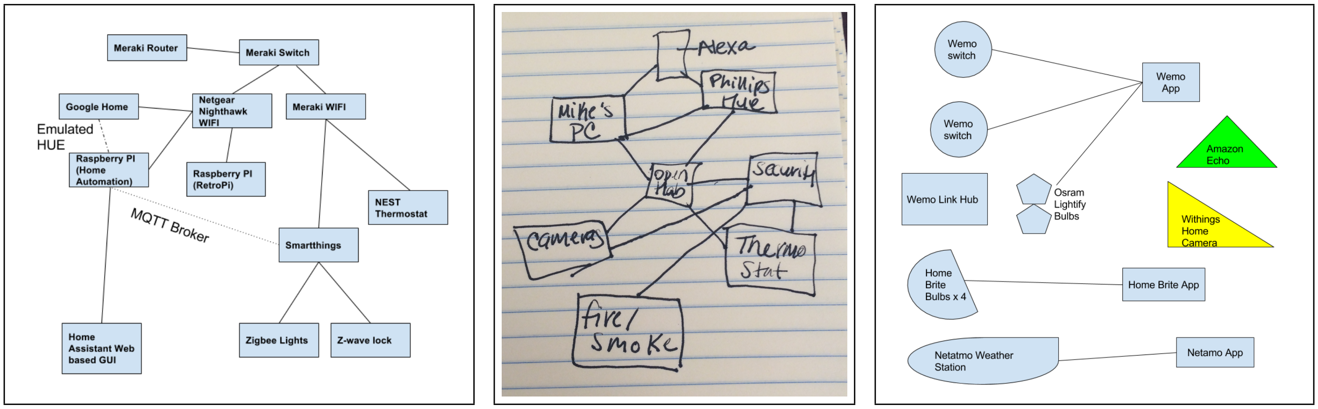


Figure 1: Participant drawings showing examples of (a) advanced (from participant P2), (b) intermediate (from participant P14), and (c) limited technical mental models (from participant P3). P2's diagram (a) shows how they used network segmentation to separate their smart home devices from their other computers. P14's diagram (b) does not represent the network topology, but rather links in functionality. In P3's diagram (c), lines are drawn between devices and their associated apps, but no technical details are captured. This diagram was edited for clarity, removing only text describing the functions of the devices.

5.3 Smart Home Technology Mental Models

Based in part on prior work linking limited technical mental models with limited adoption of security tools and incomplete threat models [34, 49], we sought to understand participants' general mental models about their smart homes before diving into security specific questions. We categorized the sophistication of participants' mental models based on both their drawings and their verbal explanation of their smart home system. Our analysis was based on codes for whether the participant demonstrated an understanding of specific technical elements of their smart home, which we describe below.

Participants with the most advanced mental models had a highly technical level of understanding of their smart home system, and were able to represent the network topology, including wireless protocols, hubs, routers, and sometimes the role of cloud servers. One example of this is P2, who was able to produce an accurate network diagram (see Figure 1a), and raised concerns about how commands traveling to the cloud affect latency. Participants in this category generally had a background in IT or computer science.

Participants with an intermediate level mental model had some sense of which devices in their home communicate with each other, but without a deep understanding of how. These users were typically capable users of technology, but did not have technical training. One participant in this category (P14) diagrammed functional relationships between devices in her home (see Figure 1b), such as between the Amazon Echo and Philips Hue lights, but did not capture the role of the cloud or their wireless router.

The last category encompasses participants who had a limited understanding of smart home technology in general, and indicated no awareness of technical details, like their network or the cloud. When we prompted to draw a diagram of his smart home system, P11 drew the physical layout of his home, and the locations of the devices, but did not illustrate how the lights and the Echo communicated with each other. Another example was P3 — in her diagram, each device had a line drawn to a shape representing the smartphone app associated with the device (see Figure 1c).

Asset	Concerned	Mentioned but not concerned
Physical security	11/15	1/15
Audio logs	4/15	4/15
General home privacy	5/15	1/15
Behavior/presence logs	2/15	2/15
Personally identifiable info	2/15	1/15
Bandwidth	1/15	0/15
Money	1/15	0/15
No identified assets	1/15	

Table 3: Assets identified by participants.

As we will see in the next section, the sophistication of a participant's technical mental model often affects the sophistication of the resulting threat model.

5.4 Smart Home Threat Models

We now turn to a core component of our study: participant threat models and security/privacy concerns (or lack thereof).

Overall, we found that participant threat models were sparse. Participants mentioned a diverse set of potential security and privacy issues, but few concrete concerns were articulated by a majority of participants. Moreover, participants were sometimes aware of potential issues but were explicitly not concerned about them. Thus, we coded threat model themes as "mentioned", "not mentioned", and "mentioned but not concerned". We summarize participant threat models in Tables 3-6, organized into assets, adversaries, vulnerabilities, and threats that came up during the interviews.

Assets. The most common asset identified by participants was physical security. This theme arose among participants who used security cameras or other security systems, or participants who mentioned concerns about door locks, which control physical access to the home.

> Most of the switches and bulbs are used to control the lightning in the home, for security purposes... The cameras are used for security, to be able to monitor the doors when we are not home, and the dog when we are away from home briefly. (P3)

Adversary	Concerned	Mentioned but not concerned
Unspecified bad actors	9/15	0/15
Company	1/15	8/15
Government	2/15	2/15
Owner of smart home devices	1/15	1/15
3rd party automation authors	1/15	0/15
No identified adversaries	2/15	

Table 4: Adversaries identified by participants.

Concern about physical security is perhaps natural, as devices like door locks and security cameras exist expressly for that purpose. Indeed, one participant cited these concerns as a reason to be more concerned about security risks with a smart home than with a laptop or phone:

> For the home, it's definitely something that I'm worried about because I don't want someone accessing the lock, or knowing the motion sensor data, or when we're home. On your phone, you have a degree of control — you can encrypt your phone, you can set up proper security — PIN and locks and stuff... I'm more concerned about my home than my phone, because if I lose my phone I can remotely wipe it. (P6)

Meanwhile, other risks with smart home devices occur as a side effect, such as privacy violations. Many participants acknowledged privacy could be an asset, particularly in the form of audio or behavior logs. However, half of these participants were not especially concerned about privacy risks:

> It's not like I openly admit to anything ridiculous that would incriminate me. And even if I did, no one's going to hear it, because Amazon doesn't release audio logs... That doesn't bother me, I guess — some people, it freaks them out, but it's not a big deal. It's just part of big data. They're just trying to gather data for advertising purposes, whatever floats their boat. (P5)

Other, less commonly identified assets that might be affected by security or privacy risks included bandwidth, money, or personally-identifiable information (PII). In one of our exploratory interviews, we heard an anecdote in which someone set up a custom smart sprinkler system which, due to an incorrect trigger, accidentally watered the lawn for a week and led to a significant water bill. Though this case was accidental, it could also be a compelling target for an attack.

Notably, no participant identified availability of device functionality as an asset that might be attacked. Although several participants voiced concerns about reliability (see Section 5.7), none connected this concern to security risks (rather identifying non-malicious network or power failures).

Adversaries. In general, when participants speculated about potential attacks on their smart home, they did not articulate specific adversaries in those scenarios (often referring to adversaries as "someone").

The most frequently identified potential adversaries were the companies that manufactured their smart home devices and that received data from those devices in the cloud. However, almost all participants who acknowledged this sort of behavior from companies were not concerned, and trusted the companies to protect their privacy. For example:

> In terms of the smart home stuff in particular, we are dealing with Amazon, we are dealing with big companies that are probably not totally irresponsible about privacy and security. (P11)

A few other participants mentioned the government as an adversary. However, they seemed to consider this concern only in the abstract sense, not providing many specifics on actions the government would take. For example, only one of the participants who mentioned the government also mentioned the murder case where law enforcement is requesting that Amazon turn over recorded audio data from an Echo device [3]. Less specifically, participants voiced general concerns about the government's surveillance capabilities and the current political climate (circa February 2017). For example:

> I am beefing up operational security in a big way, because I have spoke publicly against fascism, and I work in a publicly funded institution, I expected to be targeted at some point. (P1)

Other participants were aware of the government's potential surveillance capabilities but not overly concerned:

> I haven't changed any of my behavior in the house. If the FBI/CIA actually ever gets a recording of what's going into my Echo, they'll probably just think I'm a weirdo. (P8)

Participants had few concerns about the developers of smart home applications or custom automations as adversaries. P12 noted that the custom automations for the Vera hub were simple enough that he could read and understand it.

> Oh no, [the code] is so plain language. The only code they're writing for me is conditional commands. To turn on all the lights, I do that all myself, that's a standard scene... It's just the two tier deep programming [sic] that I've gotten their help with. And it's pretty obvious, the code they've written, I've saved it in a text file, it's you know, less than 30 characters. It's pretty obvious it's only pointing — it's like COBOL. (P12)

This lack of concern represents a gap to the threat models of security experts, who often explicitly include app developers as potential adversaries in their threat models and attempt to curtail the default capabilities of applications (e.g., [27]).

Finally, some participants were concerned or encountered issues with other residents in or visitors to the home; we discuss these issues in Section 5.6 below.

Vulnerabilities. Participants identified few concrete vulnerabilities that might lead to a security or privacy compromise, and no potential vulnerability was mentioned by a majority of participants (see Table 5). In general, we found that participants with different levels of technical knowledge identified different types of vulnerabilities in the threat model. For example, only participants with a more technically accurate mental model mentioned lack of transport level security (HTTPS) as a vulnerability.

Vulnerabilities	Concerned	Mentioned but not concerned
Data at risk in the cloud	1/15	5/15
Weak passwords	5/15	0/15
Lack of transport level security	4/15	0/15
Insecure devices	4/15	0/15
Malicious devices	3/15	0/15
Unsecured Wi-Fi network	2/15	0/15
Devices can be unpaired	1/15	0/15
No identified vulnerabilities	3/15	

Table 5: Vulnerabilities identified by participants.

Threats	Concerned	Mentioned but not concerned
Continuous audio/video recording	3/15	5/15
Data collection and mining	1/15	5/15
Adversarial remote control	4/15	1/15
Network attack on local devices	3/15	1/15
Spying by other user in home	3/15	0/15
Account/password hacking	2/15	0/15
Network mapping by mal. devices	1/15	0/15
Re-pair device with attacker's hub	1/15	0/15
No identified threats	1/15	

Table 6: Threats identified by participants.

> A lot of stuff is just totally unencrypted. Some of it is encrypted, a lot of it doesn't validate SSL certs. ... Even today, there's a lot of use of weak encryption ciphers. Yeah, it's pretty awful. (P1)

Meanwhile, participants with a less sophisticated mental model were more concerned about weak passwords and unsecured Wi-Fi networks, which are vulnerabilities that are not specific to the smart home context.

> People are concerned that someone could check into their camera or their lights... I guess they're not smart enough to know that they can't do that if they don't get your password. (P3)

Some participants mentioned concerns about malicious or vulnerable devices, either specifically (e.g., P8 was aware of Foscam web cam vulnerabilities [18]) or more generically:

> I think the biggest thing is just the amount of questionable things that have happened within the IoT space from some of the up and coming companies. That has me questioning what they can and can't do... I've just heard horror stories from some of the smaller companies. (P2)

Threats. As with vulnerabilities, there was not a particular threat or attack that a majority of participants were concerned about (see Table 6). While many acknowledged that companies or other adversaries *could* record and store private data, like audio/video feeds and behavioral logs, again we found that most were not concerned about it.

Again, we saw that participants with more advanced mental models voiced more concrete and technical threats, such as network attacks and network mapping. For example, P10 identified a specific threat: that an adversary with physical access to the home could un-pair a device from the user's hub, and re-pair it with their own hub.

Reasons for Lack of Concern. Even when participants were aware of security and privacy issues, they were often not actively concerned about them, voicing several reasons.

One reason for lack of concern, discussed above, is explicit trust in companies handling user data, such as Amazon.

Some participants were not concerned about attacks because they did not consider themselves a worthwhile target (notably, not considering untargeted attacks like widespread DDoS):

> I read some stuff about Hue bulbs being hacked, but I live in a small town. No one is going to pull up to my house and do any of that stuff. (P7)

Some believed they have nothing to hide, a perception that other researchers have reported for online behavior [15]. Others believed that they had taken sufficient steps to secure their systems, such as with strong passwords, so they did not need to worry further about security. For example:

> I also know many, many people who have such powerfully weak passwords, that if someone were driving around trying to get into someone's stuff, they would get into someone's stuff with weak passwords, and not into mine. (P3)

> I see the ability for devices to be manipulated if not secured properly, but from what I've read it seems like you can lock your system down pretty well, by just having a secure network and backup options. (P4)

Seven participants explicitly identified a tradeoff, requiring that one accepts security or privacy risks in exchange for the functionality and convenience of a smart home. For example:

> ...your data's going somewhere, and it comes down to who you are going to trust with it. You can trust it with Amazon, who has a record of everything you have spoken to your Echo, or are you gonna trust it with Google, who has access to your email, your map search history, your web search history? It depends on who you think is gonna do what with your data. ... It's a tradeoff of these free services — you're getting Gmail for free, but you're letting them run ads. (P6)

> I think our security is so compromised in so many different ways and I'm broadly speaking willing to accept some of the benefits of having these system understand my life — targeted advertising and various other conveniences. (P11)

5.5 Mitigation Strategies

Here, we consider approaches participants took to mitigate their security and privacy concerns. Mitigation strategies varied greatly, with no single strategy shared by more than five participants, suggesting that best practices for end user smart home security have not become standard.

Technical Mitigations. Two participants intentionally kept their smart home devices on a separate Wi-Fi network from other home electronics, perhaps concerned about attacks by compromised smart home devices on other electronics, which may have more valuable data. These participants also blocked certain traffic from their devices: P1 blocked all unencrypted traffic, and P2 prevented their SmartThings hub from communicating with cloud servers, instead using an MQTT broker to control it from a local server.

Some participants attempted to mitigate password and Wi-Fi security related concerns with best practices, presumably learned from more traditional computing contexts:

> I don't have any security concerns because I feel fairly confident that the — I know that my passwords for all those accounts are very secure. (P3)

A few participants, with more technical backgrounds, desired additional security or privacy features on their devices, such as better use of HTTPS, or more granular permissions on the sensors on devices. P1 in particular wanted to be able to switch off the microphone on a Nest thermostat device. Only two participants mentioned deleting camera recordings or other logs of behavior to protect their privacy.

In some cases, participants used mitigation strategies with unclear benefits, suggesting limited underlying technical knowledge. For example, P7 only used Z-Wave smart home devices for security reasons, and when asked why, he said:

> I don't really remember. There was an article I was reading about... it was when I started out like two years ago that I researched it and I got these things in my head... I don't remember the specifics. I'm not an expert on any of this stuff. I try to do my research, but I have to take other people's opinion at face value. (P7)

Non-Technical Mitigations. A possible strategy for mitigating privacy risks in smart homes is simply altering one's behavior around those devices. For example, one might avoid saying certain things around the Echo or doing certain things in front of cameras. However, when asked explicitly about such behavior changes, nine participants explicitly mentioned that they did not change their behavior at all. Others only mentioned changing their behavior in theory:

> If I was to do something illegal I wouldn't do it in the room that has the Alexa and the camera in it. I would probably also turn off my cell phone, because... you don't know. I generally don't feel concerned because I'm not currently up to anything that is so private that it can't be stored in Amazon's temporary voice audio recording database. (P13)

Several participants made choices about where to place devices, or when those devices were enabled, for privacy reasons:

> We choose not to face [the camera towards] any interior portions. I do have a camera that's easy to set up, and when we go out of town for a couple of days, I'll just plug it in and it faces the interior, but never when we're actually home. (P1)

> With the camera I have in the house, I do have it plugged it into [a smart] powerstrip. So I don't really need that on when I'm there. So that's one thing that I guess we did do something little different, just have the camera come on when we're away. (P7)

5.6 Multi-User Interactions

We now turn to concerns and issues related to incidental users of the smart home, who were not primarily involved in selecting or automating devices. Three participants were incidental users, and we also asked primary users about disagreements with or concerns of incidental users.

Differences in Mental and Threat Models. We found that in general, incidental users of smart homes have simpler mental models, less awareness of security/privacy issues, and weaker threat models. This is perhaps natural; the person who wants to set up a smart home is likely more enthusiastic and curious about researching the technology, while the other resident(s) might simply tolerate their smart home "hobby".

For example, P14 lives in a fairly complex smart home set up by her husband (who is seemingly tech-savvy, as their OpenHAB hub requires programming skills). However, P14's mental model of their smart home is incomplete (see her drawing in Figure 1b), and she deferred most of the worrying about security to her husband. When asked specifically if they had security or privacy concerns, she said:

> It is something we joke about, but he's assured me that no one's going to be able to hack into it. I don't know if I believe that. (P14)

Differences in Access. Additionally, we found that incidental users do not always have full access to the smart home. Often they do not have the proper apps installed to control the home, either because the devices can be controlled without the app, using an Amazon Echo or Google Home, or these users were simply not interested in playing with the app and setting up automations on their own.

Differences in Power and Control. One consequence of non-primary users having less access and less interest in smart homes is that it leads to situations where the primary user may have — intentionally or unintentionally — more power over the other residents of the home. For example, we observed three such cases in our interviews.

Case 1: Restricted Access. P5 did not give their spouse access to the thermostat, because they wanted to keep it at a certain temperature to save power:

> I locked down my thermostat from [my wife] specifically. Because she complains that it is hot all the time, and I'm like, "Just turn on the fan, just turn on the ceiling fan and stand under it, and you'll be good," because it costs money. (P5)

Case 2: Audio/Video Surveillance. P13 lives in a house where the smart home setup was provided by the landlord. In particular, they had an Amazon Echo, a Nest surveillance camera, and Philips Hue lights. The landlord, being the owner of the devices, had accounts associated with these devices. That gave the landlord access to transcriptions and recordings of voice queries to the Echo, and could receive notifications from the security camera. The landlord accessed private data in at least one instance:

> We threw a party and didn't tell the woman who coordinates our house, and someone unplugged the Nest camera in the kitchen because they wanted to recharge their phone... and when it is unplugged, it automatically sends an email to whoever's account is associated with the camera, and it has a photo of the last thing the camera saw. So we were throwing this huge party, and it sent a photo of the kitchen..., and so the coordinator got the email and it was like "Your camera was unplugged, this is the last thing the camera saw!"... She wasn't mad! They were excited that we were having a party. (P13)

In this instance, there were no negative consequences, nor was P13 particularly concerned about their landlord's access to the smart home, but other situations may not be as benign.

Case 3: Behavioral Surveillance. P2 has an extensive smart home setup, and mentioned using the smart lock to find out when his wife and children arrived at home. In addition, he had custom software to detect when devices were on the network, which also indicates who is at home. When asked about disagreements with his wife, P2 said:

> My wife hates the aspect that I know when her device comes or goes on the local LAN, which obviously creates an audit log, so to speak, of when she's at home. She's now chiming in, that's the reason her phone doesn't connect to the Wi-Fi anymore, so I can't track her. (P2)

In this case, P2's surveillance does not appear to have been malicious but rather a result of his experimentation with the smart home — but again, other situations may be more dangerous (e.g., domestic abuse [36]).

Trolling. On a more lighthearted note, participants identified several instances of "trolling" among residents or guests in a smart home. Though these examples are not malicious and were not poorly received, they also highlight potential tensions that may arise between multiple users. For example:

> I had my family here over the weekend, and they have an Echo as well... They said "Hey Alexa, put poop on my shopping list" and then they said "Hey Alexa, order that", and of course it said "Are you sure?" and they let me say no. (P5)

5.7 Non-Security and Privacy Concerns

Finally, participants often cited non-security and privacy related concerns that influenced how they set up their smart home system. These concerns can be at odds with security and privacy, and researchers or platform designers focused on addressing security and privacy issues must consider these other constraints as well.

Reliability. Eight participants expressed concern about their home's resilience to network and power failures. In the event that their devices and hubs could no longer connect to the internet, participants wanted their devices to continue to function as normal, including their automations. For these participants, the ability to run automations locally was a deciding factor on which hub they decided to buy.

> If... you're trying to do something, and it doesn't work because the internet is down, that's really annoying... your wall switches should still work, your automations might not work, but simple stuff that doesn't require the internet to process things should still work. (P6)

As discussed above, despite this concern about reliability, no participant considered it in the context of security, i.e., no one mentioned that availability could be impacted maliciously. We also observe that maliciously induced failures could be leveraged for other attacks, e.g., to access door locks, although no participant voiced such a concern either.

Interoperability. Six participants mentioned that they want their devices to be interoperable, i.e., compatible with the rest of their smart home system. These participants would like their devices to be controllable by their hub, by their Echo/Google Home, and/or by a single, centralized app. As discussed above, in several cases participants installed third-party applications, such as IFTTT or Stringify, for the sole purpose of making devices interoperable. Such ad hoc connections potentially introduce new security vulnerabilites by expanding the attack surface of their system.

Cost. For some participants, a more prominent barrier to adoption was device cost. For example, this led P10 to cobble together his own "smart" sprinkler system rather than buying an existing smart device, increasing the risk of user error and potentially opening the door for security vulnerabilities.

5.8 Results Summary

Before stepping back in Section 6, we summarize our key findings from interviews with smart home end users:

- Participants have varied and sparse threat models, and do not share a common set of concerns or mitigations.
- Participants' threat models often depend on the sophistication of their technical mental models.
- Reasons for lack of concern about security/privacy issues include not feeling personally targeted, trusting potentially adversarial actors (like companies or governments), and believing their existing mitigation strategies to be sufficient.
- Concerns of security experts about smart homes, such as insecure or malicious devices, company data collection, attacks on device availability, or malicious or buggy third-party apps, were generally not shared by participants.
- Homes with multiple users pose unique security and privacy challenges, especially when the primary user has greater knowledge and control of the system than incidental users.
- Participants make smart home technology choices based on requirements that may conflict with security and privacy, including cost and interoperability.

6. DISCUSSION

We now step back to reflect on lessons from our findings, develop recommendation, and discuss study limitations.

6.1 Lessons

Incomplete mental models lead to gaps in threat models and security behaviors. Echoing prior work on mental models and security [34, 49], we found that participants with more sophisticated mental models had more advanced threat models that identified risks unique to smart homes, and were able to take specific precautions to address these risks, such as blocking unencrypted traffic from their smart home devices. On the other hand, participants with less sophisticated mental models did not identify smart home-specific vulnerabilities and threats, and often based their mitigation strategies on best practices from other technologies, like using strong passwords, or adopted ad hoc strategies with unclear benefits, like avoiding using non-Zigbee devices.

The absence of common threat model elements and mitigation strategies suggests that best practices for smart home security have yet to be developed. This gap makes it difficult for users without a deep technical understanding of the technology to make informed security decisions.

Participants were more about physical security issues than privacy issues. The physical security of the home was a common concern voiced by our participants. This concern was expressed in two ways: either concern about attackers compromising security-critical devices like smart locks, or using the smart home to enhance their home security, with light timers and cameras. Brush et al. [9] found similar concerns: remote access to locks and cameras is important but creates a security risk.

However, most participants were unconcerned about privacy issues with their smart homes, despite having at least a cursory awareness. A possible explanation for this result could be that devices like door locks have security as their *primary* purpose, so a security failure would be equivalent to a functionality failure. By contrast, privacy risks with other devices are side-effects of their intended purpose, (e.g., privacy risks due to the Echo's ability to record audio).

This result could also be explained in part by our participant group: smart home users. These users have already chosen to set up a smart home (or had one set up for them); we did not hear from people who chose *not* to install a smart home due to security and privacy concerns. We discuss this limitation further in Section 6.3 below.

Mismatch between awareness and power of smart home administrator and other residents. In addition to replicating findings about the primary/incidental user dynamic from previous studies of end users of smart homes [9, 20, 41, 40], which found that most households have one user who is more active about researching, purchasing, and setting up smart home devices, our findings suggest that incidental users of smart homes may be less tech-savvy and/or less informed or aware about potential security and privacy issues. These discrepancies can lead to a power imbalance in a home, with the primary user in a position to (maliciously or not) spy on other residents or limit their control of the home.

A key observation here is that while the people who set up smart homes, particularly early adopters, often treat the technology as a personal hobby, smart homes are fundamentally *not* personal technologies. As a result, any security and privacy (or other) decisions made by the primary user directly affects other residents and visitor. If the primary and incidental users share a threat model, this interaction can be positive; however, if they do not agree on concerns or, worse, the primary user is adversarial (e.g., abusive) towards the incidental users, dangerous situations can arise.

Flexible end user programming limits usefulness of third-party applications. We found that users make limited use of third-party apps, e.g., on Samsung SmartThings. Instead, they more frequently use end-user programming interfaces (e.g., to set custom automation rules) or directly write scripts. When third-party apps were used, it was often to connect other ecosystems to the platform, e.g., to enable Amazon Echo based voice control of Samsung devices. This finding begs the question: Why? Are packaged apps not sufficiently flexible for diverse home environments (unlike on more homogenous smartphones)? Do they not yet provide sufficiently compelling functionality? Future work may shed light on these questions; in the meantime, the app platforms may not be the most critical place for the security community to focus its efforts — as otherwise seems natural, given the wealth of work on smartphone app platforms.

6.2 Recommendations and Future Work

We develop recommendations for the designers of smart home platforms and devices, as well as for future research.

UI/UX for User Awareness and Control. By improving users' technology mental models, we can also improve the accuracy of their threat models, enabling conscious decisions about whether to mitigate or ignore privacy or security risks. A possible strategy is to surface more information to users about what devices are doing — e.g., by providing usable auditing features in the associated phone apps, or by including physical indicators on devices (e.g., recording lights). Consolvo et al. used similar techniques for surfacing information leakage over unencrypted Wi-Fi networks [13]. Such indicators are already common for cameras and microphones, but may not be noticed by users performing unrelated tasks [46] — thus, future work must study how to design effective indicators in the smart home context, where users are often not directly interacting with devices.

Similarly, user control can be enhanced by ensuring that users can interact with devices physically, not only through apps. This can improve multi-user interactions and can help mitigate potential impacts of network failures, cloud outages, and phone or app problems. Indeed, several participants explicitly mentioned the need for physical switches.

> If the system is shut off, your wall switches should still work, your automations might not work, but simple stuff that doesn't require the internet to process things, it should still work. (P6)

Design Consciously for Multiple Users. In many of today's smart home platforms, support for multiple users is overlooked, and platform designers seem not yet to have deeply considered the potential risks among users in the same home. For example, users of SmartThings can easily monitor other users, and the Echo allows access to audio logs. From our interviews, we also heard about cases in which incidental users were intentionally or unintentionally denied access to smart home controls. As smart homes become more prevalent, similar issues may arise with guests.

Thus, future smart home platforms must take into account multi-user interactions and the potential power imbalance between the primary user and incidental (and often less tech-savvy) users. In addition to the need to support multiple distinct user accounts, usability and discoverability of features are critical for secondary, less technical users.

The user control and awareness recommendations we make above can also help improve the multi-user experience. For example, if devices can all be controlled with physical switches, then all residents are guaranteed the ability to control that device. Similarly, physical recording indicators and other usable audit logs can help improve awareness of incidental users. We encourage future research to further study both the dynamics of multi-user smart homes as well as evaluate potential designs to mitigate these issues.

Reputation Systems for Smart Home Options. Not all users can (or should) become technical experts. Instead, external guidance may be required to help users make informed decisions about which products have stronger security and privacy properties. For smartphones, centralized app stores provide app reputation information, and prior work [25] has shown that users use these reviews to make decisions about which applications to install, including for security and privacy reasons. Similarly, the Electronic Frontier Foundation's secure messaging scorecard [23] aims to inform users about the security properties of different messaging options. The smart home ecosystem is much more heterogeneous, and there are no well-known centralized resources for security or privacy sensitive users to inform themselves. The recent news that Consumer Reports will begin evaluating products for security and privacy [14] is a step in the right direction.

Develop Standard Best Practices for End Users. As discussed above, we found that participants often adopted best security practices not specific to smart homes, such as strong password and Wi-Fi security. However, these practices do not cover many of the security and privacy issues unique to smart homes, and security experts must develop — and communicate effectively to end users — an updated set of best practices for smart home contexts. For example, one such recommendation might be unplugging or muting recording devices when they are not needed or during sensitive conversations, or alerting guests to their presence.

Design for Secure and Robust Interoperability. Interoperability between devices and smart home ecosystems (e.g., Amazon and SmartThings) was important to many participants, who often installed third-party apps or built custom solutions to connect different pieces of their smart home. Since security issues often arise at the boundaries between components, these user-created interoperability links (likely different across individual smart homes) may present future points of weakness. Security researchers should study these integrations, and smart home platform and devices designers should explicitly design for robust interoperability.

Minimize Tradeoffs for Security and Privacy. Many participants identified a tradeoff between security and privacy with functionality and convenience, in some cases sounding resigned to it. We challenge smart home designers and researchers to present a better tradeoff. For example, certain technical design choices can reduce risks without significantly impacting functionality, like not requiring the cloud to run automations. (Indeed, SmartThings initially only supported running apps in their cloud, but now supports apps on the local hub [51] — although perhaps for reliability rather than security reasons.) By minimizing these tradeoffs when possible, we can remove the decision-making burden from users and enable adoption of smart home technologies by people who are not willing to make the tradeoffs required today.

6.3 Limitations

Finally, we reflect on several limitations of this work. First, we only interviewed participants living in smart homes, not people who chose not to install them for security or privacy (or other) reasons. Future work should study this deliberate non-user population, as they may have more pronounced concerns that hindered adoption in the first place. Here, our focus was on participants who could speak to concrete, rather than hypothetical, smart home experiences.

Second, our sample skews towards primary users and smart home enthusiasts, despite our efforts to recruit more passive users of smart homes. This is likely in part due to self-selection bias among people drawn to participate in our study, and because we recruited from smart home-focused online communities. Nevertheless, these participants' accounts of other residents, as well as the interviews with three non-primary users, help shed light on this class of end users.

Third, smart home technology is new and still developing — commercial platforms targeted at non-technical consumers, like Samsung SmartThings or Amazon Echo, are recent developments. Thus, our participants are among the earliest adopters; they may be more willing to choose convenience over security or privacy, or be generally more tech-savvy, than non-adopters. As smart homes become more widespread, the makeup of the user base will shift, and future work should consider these changes. Meanwhile, our findings already shed light on issues that will arise and become more complex, e.g., around multi-user scenarios, as adoption increases.

Finally, this qualitative study, by its nature, does not produce quantitative conclusions, e.g., about the prevalence of certain concerns or lack thereof. This work lays the groundwork for future quantitative studies to investigate such questions.

7. CONCLUSION

Consumer smart home technologies are becoming increasingly prevalent. Alongside the convenience offered by these technologies, they raise new security and privacy risks. Though researchers have begun studying these technologies themselves, there has been little study of the end users of modern smart homes: what are their mental models, security and privacy concerns, mitigation strategies, and how does the presence of multiple users compound these issues? We sought to answer these questions in our work, conducting in-depth interviews with fifteen participants (twelve smart home administrators and three other residents). Our findings shed light on their mental models and security concerns (or lack thereof) — for example, revealing incomplete threat models and ad hoc mitigation strategies based on best practices for older technologies — and highlight potential tensions between multiple smart home users. These findings lay the groundwork for continued study of smart home end users as the technologies develop further and see increased adoption, and we provide recommendations to smart home technology designers and researchers for where to focus future efforts. For example, we highlight the need to help shape user mental models, consciously design for multiple users, and design for security and privacy alongside key features valued by users (e.g., interoperability and remote access).

Acknowledgements

We are especially grateful to our user study participants, as well as our pilot study participants, Camille Cobb and Alex Takakuwa. We thank our anonymous reviewers and our shepherd, Blase Ur, for helpful feedback on an earlier version. We also thank Luis Ceze for useful conversations about smart home setups, as well as Yoshi Kohno, Kiron Lebeck, Lucy Simko, and Anna Kornfeld Simpson for reviewing an earlier draft. This work was supported in part by the National Science Foundation under Awards CNS-1513584 and CNS-1565252, and by a Hacherl Endowed Fellowship.

8. REFERENCES

[1] Vera smarter home control. Accessed March 7, 2017, Online at http://getvera.com/.

[2] Wink hub. Accessed March 7, 2017, Online at https://www.wink.com/products/wink-hub/.

[3] Amazon shares data with Arkansas prosecutor in murder case, Mar. 2017. Accessed March 6, 2017, Online at http://bigstory.ap.org/article/1110e4449c3f4191909e4010da935056/amazon-shares-data-arkansas-prosecutor-murder-case.

[4] Amazon Echo. Accessed March 7, 2017, Online at https://www.amazon.com/Amazon-Echo-Bluetooth-Speaker-with-WiFi-Alexa/dp/B00X4WHP5E.

[5] Apple HomeKit. Accessed March 7, 2017, Online at https://www.apple.com/ios/home/.

[6] O. Arias, J. Wurm, K. Hoang, and Y. Jin. Privacy and security in Internet of Things and wearable devices. *IEEE Transactions on Multi-Scale Computing Systems*, 1(2):99–109, Nov. 2016. DOI 10.1109/TMSCS.2015.2498605.

[7] K. Baxter, C. Courage, and K. Caine. *Understanding Your Users: A Practical Guide to User Research Methods*. Morgan Kaufmann, second edition, 2015.

[8] N. Bilton. Nest thermostat glitch leaves users in the cold. The New York Times, Jan. 2016. Accessed March 7, 2017, Online at https://www.nytimes.com/2016/01/14/fashion/nest-thermostat-glitch-battery-dies-software-freeze.html.

[9] A. J. Brush, B. Lee, R. Mahajan, and S. Agarwal. Home automation in the wild: Challenges and opportunities. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 2011. DOI 10.1145/1978942.1979249.

[10] J. bum Woo and Y. kyung Lim. User experience in do-it-yourself-style smart homes. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2015. DOI 10.1145/2750858.2806063.

[11] D. J. Butler, J. Huang, F. Roesner, and M. Cakmak. The privacy-utility tradeoff for remotely teleoperated robots. In *Proceedings of the Annual ACM/IEEE International Conference on Human-Robot Interaction*, 2015. DOI 10.1145/2696454.2696484.

[12] E. K. Choe, S. Consolvo, J. Jung, B. L. Harrison, S. N. Patel, and J. A. Kientz. Investigating receptiveness to sensing and inference in the home using sensor proxies. In *Proceedings of the International Conference on Ubiquitous Computing (UbiComp)*, 2012. DOI 10.1145/2370216.2370226.

[13] S. Consolvo, J. Jung, B. Greenstein, P. Powledge, G. Maganis, and D. Avrahami. The Wi-Fi privacy ticker: improving awareness & control of personal information exposure on Wi-Fi. In *Proceedings of the International Conference on Ubiquitous Computing (UbiComp)*, 2010. DOI 10.1145/1864349.1864398.

[14] Consumer Reports. Consumer Reports to begin evaluating products, services for privacy and data security. Accessed March 7, 2017, Online at http://www.consumerreports.org/privacy/consumer-reports-to-begin-evaluating-products-services-for-privacy-and-data-security/.

[15] G. J. Conti and E. Sobiesk. An honest man has nothing to fear: User perceptions on web-based information disclosure. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 2007. DOI 10.1145/1280680.1280695.

[16] K. L. Courtney, G. Demeris, M. Rantz, and M. Skubic. Needing smart home technologies: The perspectives of older adults in continuing care retirement communities. *Informatics in Primary Care*, 16(3):195–201, 2008.

[17] A. Cui and S. J. Stolfo. A quantitative analysis of the insecurity of embedded network devices: Results of a wide-area scan. In *Proceedings of the Annual Computer Security Applications Conference*, pages 97–106. ACM, 2010. DOI 10.1145/1920261.1920276.

[18] CVE. CVE-2013-2560, 2013. Accessed March 7, 2017, Online at https://cve.mitre.org/cgi-bin/cvename.cgi?name=CVE-2013-2560.

[19] N. Dell, V. Vaidyanathan, I. Medhi, E. Cutrell, and W. Thies. "Yours is better!": Participant response bias in HCI. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 2012. DOI 10.1145/2207676.2208589.

[20] A. Demeure, S. Caffiau, E. Elias, and C. Roux. Building and using home automation systems: A field study. In *Proceedings of International Symposium on End User Development (IS-EUD)*, 2015. DOI 10.1007/978-3-319-18425-8_9.

[21] T. Denning, T. Kohno, and H. M. Levy. Computer security and the modern home. *Communications of the ACM*, 56(1):94, Jan. 2013. DOI 10.1145/2398356.2398377.

[22] T. Denning, C. Matuszek, K. Koscher, J. R. Smith, and T. Kohno. A spotlight on security and privacy risks with future household robots: Attacks and lessons. In *Proceedings of the International Conference on Ubiquitous Computing (UbiComp)*, 2009. DOI 10.1145/1620545.1620564.

[23] Electronic Frontier Foundation. Secure messaging scorecard. Accessed March 7, 2017, Online at https://www.eff.org/secure-messaging-scorecard.

[24] N. Feamster, S. Grover, and R. Ensafi. Who will secure the Internet of Things?, Jan. 2016. Accessed March 7, 2017, Online at https://freedom-to-tinker.com/2016/01/19/who-will-secure-the-internet-of-things/.

[25] A. P. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D. Wagner. Android permissions: User attention, comprehension, and behavior. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 2012. DOI 10.1145/2335356.2335360.

[26] E. Fernandes, J. Jung, and A. Prakash. Security analysis of emerging smart home applications. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*, 2016. DOI 10.1109/SP.2016.44.

[27] E. Fernandes, J. Paupore, A. Rahmati, D. Simionato, M. Conti, and A. Prakash. FlowFence - Practical data protection for emerging IoT application frameworks. In *Proceedings of the USENIX Security Symposium (USENIX Security)*, 2016.

[28] J. L. Fleiss, B. Levin, and M. C. Paik. *Statistical Methods for Rates and Proportions*. John Wiley & Sons, 3 edition, 2003.

[29] Google Home. Accessed March 7, 2017, Online at https://madeby.google.com/home/.

[30] J. Granjal, E. Monteiro, and J. Sa Silva. Security for the Internet of Things: A survey of existing protocols and open research issues. *IEEE Communications Surveys and Tutorials*, 17(3):1294–1312, 2015. DOI 10.1109/COMST.2015.2388550.

[31] K. Hill. 'Baby monitor hack' could happen to 40,000 other Foscam users, Aug. 2013. Accessed March 4, 2017, Online at https://www.forbes.com/sites/kashmirhill/2013/08/27/baby-monitor-hack-could-happen-to-40000-other-foscam-users/#40c00e3a58b5.

[32] G. Ho, D. Leung, P. Mishra, A. Hosseini, and D. Song. Smart locks: Lessons for securing commodity Internet of Things devices. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, 2016. DOI 10.1145/2897845.2897886.

[33] D. Jonassen and Y. H. Cho. Externalizing mental models with mindtools. In *Understanding models for learning and instruction*, pages 145–159. Springer, 2008.

[34] R. Kang, L. A. Dabbish, N. Fruchter, and S. B. Kiesler. "My data just goes everywhere:" User mental models of the internet and implications for privacy and security. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 2015. Online at https://www.usenix.org/system/files/conference/soups2015/soups15-paper-kang.pdf.

[35] F. Kawsar and A. J. B. Brush. Home computing unplugged: Why, where and when people use different connected devices at home. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2013. DOI 10.1145/2493432.2493494.

[36] T. Matthews, K. O'Leary, A. Turner, M. Sleeper, J. P. Woelfer, M. Shelton, C. Manthorne, E. F. Churchill, and S. Consolvo. Stories from survivors: Privacy & security practices when coping with intimate partner abuse. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 2017. DOI 10.1145/3025453.3025875.

[37] M. L. Mazurek, J. P. Arsenault, J. Bresee, N. Gupta, I. Ion, C. Johns, D. Lee, Y. Liang, J. Olsen, B. Salmon, R. Shay, K. Vaniea, L. Bauer, L. F. Cranor, G. R. Ganger, and M. K. Reiter. Access control for home data sharing: Attitudes, needs and practices. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pages 645–654, 2010. DOI 10.1145/1753326.1753421.

[38] A. McLean. IoT malware and ransomware attacks on the incline: Intel Security, Sept. 2015. Accessed March 4, 2017, Online at http://www.zdnet.com/article/iot-malware-and-ransomware-attacks-on-the-incline-intel-security/.

[39] E. McReynolds, S. Hubbard, T. Lau, A. Saraf, M. Cakmak, and F. Roesner. Toys that listen: A study of parents, children, and internet-connected toys. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 2017. DOI 10.1145/3025453.3025735.

[40] S. Mennicken, J. Hofer, A. K. Dey, and E. M. Huang. Casalendar: A temporal interface for automated homes. In *Proceedings of the Conference on Human Factors in Computing Systems: Extended Abstracts (CHI EA)*, 2014. DOI 10.1145/2559206.2581321.

[41] S. Mennicken and E. M. Huang. Hacking the natural habitat: An in-the-wild study of smart homes, their development, and the people who live in them. In *Proceedings of the International Conference on Pervasive Computing (Pervasive)*, 2012. DOI 10.1007/978-3-642-31205-2_10.

[42] P. Morgner, S. Mattejat, and Z. Benenson. All your bulbs are belong to us: Investigating the current state of security in connected lighting systems. *CoRR*, abs/1608.03732, 2016.

[43] L. H. Newman. The botnet that broke the Internet isn't going away. Wired, Dec. 2016. Accessed March 7, 2017, Online at https://www.wired.com/2016/12/botnet-broke-internet-isnt-going-away/.

[44] T. Oluwafemi, T. Kohno, S. Gupta, and S. Patel. Experimental security analyses of non-networked compact fluorescent lamps: A case study of home automation security. In *Proceedings of the Learning from Authoritative Security Experiment Results (LASER)*, 2013. Online at https://www.usenix.org/laser2013/program/oluwafemi.

[45] Philips Hue. Accessed March 7, 2017, Online at http://www2.meethue.com/en-US.

[46] 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 Conference on Human Factors in Computing Systems (CHI)*, 2015. DOI 10.1145/2702123.2702164.

[47] F. Raja, K. Hawkey, and K. Beznosov. Revealing hidden context: Improving mental models of personal firewall users. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, 2009.

[48] The home automation market by the numbers. Remotely, June 2015. Accessed March 7, 2017, Online at https://blog.remotely.com/2015/06/20/the-home-automation-market-by-the-numbers/.

[49] K. Renaud, M. Volkamer, and A. Renkema-Padmos. Why doesn't Jane protect her privacy? In *Proceedings of the International Symposium on Privacy Enhancing Technologies (PETS)*, 2014. DOI 10.1007/978-3-319-08506-7_13.

[50] E. Ronen, C. O'Flynn, A. Shamir, and A.-O. Weingarten. IoT goes nuclear: Creating a ZigBee chain reaction. Cryptology ePrint Archive, Report 2016/1047, 2016. Accessed March 7, 2017, Online at http://eprint.iacr.org/2016/1047.

[51] Samsung SmartThings. Local processing. Accessed

March 7, 2017, Online at https://support.smartthings.com/hc/en-us/articles/209979766-Local-processing.

[52] S. Shane, M. Mazzetti, and M. Rosenberg. WikiLeaks releases trove of alleged C.I.A. hacking documents. The New York Times, Mar. 2017. Accessed March 7, 2017, Online at https://www.nytimes.com/2017/03/07/world/europe/wikileaks-cia-hacking.html.

[53] S. Sicari, A. Rizzardi, L. A. Grieco, and A. Coen-Porisini. Security, privacy and trust in Internet of Things: The road ahead. *Computer Networks*, 76:146–164, Jan. 2015. DOI 10.1016/j.comnet.2014.11.008.

[54] A. K. Simpson, S. N. Patel, F. Roesner, and T. Kohno. Securing vulnerable home IoT devices with an in-hub security manager. Technical Report UW-CSE-2017-01-01, University of Washington, 2017.

[55] SmartThings. Accessed March 7, 2017, Online at https://www.smartthings.com/.

[56] Smart home. Statista Digital Market Outlook. Accessed March 4, 2017, Online at https://www.statista.com/outlook/279/109/smart-home/united-states.

[57] D. Townsend, F. Knoefel, and R. Goubran. Privacy versus autonomy: a tradeoff model for smart home monitoring technologies. In *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*, pages 4749–4752. IEEE, 2011.

[58] K. Toyama. *Geek heresy: Rescuing social change from the cult of technology.* PublicAffairs, 2015.

[59] B. Ur, J. Jung, and S. Schechter. The current state of access control for smart devices in homes. In *Workshop on Home Usable Privacy and Security (HUPS)*, 2013.

[60] B. Ur, J. Jung, and S. E. Schechter. Intruders versus intrusiveness: Teens' and parents' perspectives on home-entryway surveillance. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2014. DOI 10.1145/2632048.2632107.

[61] P. Worthy, B. Matthews, and S. Viller. Trust me: Doubts and concerns living with the Internet of Things. In *Proceedings of the ACM Conference on Doubts and Concerns Living with the Internet of Things*, June 2016. DOI 10.1145/2901790.2901890.

[62] D. J. Wu, A. Taly, A. Shankar, and D. Boneh. Privacy, discovery, and authentication for the Internet of Things. In *Proceedings of the European Symposium on Research in Computer Security (ESORICS)*, pages 301–319. Springer International Publishing, 2016. DOI 10.1007/978-3-319-45741-3_16.

[63] T. Yu, V. Sekar, S. Seshan, Y. Agarwal, and C. Xu. Handling a trillion (unfixable) flaws on a billion devices: Rethinking network security for the Internet-of-Things. In *Proceedings of the Workshop on Hot Topics in Networks (HotNets Workshop)*, 2015. DOI 10.1145/2834050.2834095.

APPENDIX

A. INTERVIEW PROTOCOL

General Questions:

1. What devices do you own?
2. Can you briefly describe what you use them for?
3. What apps or automations do you have installed?
4. Do you access devices remotely, or only when you're physically in your home?

Mental Models:
For this next part, I'd like you to draw a diagram of how all of your devices are connected together. I can either email you a link to a Google Docs drawing that we can both edit, or you can draw it on a piece of paper and send it to me.

Security and Privacy Concerns:
(Start with questions not explicitly about security or privacy:)

1. When setting up your home, did you have any hesitations about getting any devices?
2. Are there any devices you thought about getting but decided not to get? Why?
3. Are there any devices that you used to use but later deactivated?

(Move on to direct questions if they have not already started talking about security and privacy:)

1. One type of concern we're interested in is security or privacy concerns. Do you or did you have any concerns like that about your smart home? You might not have any such concerns – that's fine, and we'd like to hear about that too.

   (OR, if security/privacy have been brought up organically:)

   Do you or did you have any other security or privacy concerns that you haven't mentioned yet?
2. Have you heard about any security or privacy issues with smart homes in the news? If so, did that news concern you, or do you think those issues are a little overblown?
3. How would you compare your level of security/privacy concern about your smart home devices to your level of concern about your phone or laptop computer?
4. *(For Echo/security camera users:)* Do you ever look at the audio/video logs of your Echo/camera?

Mitigation Strategies:

1. Thinking specifically about security and privacy concerns, have those concerns caused you to change any of your behaviors?
   (a) For example, do you act differently in your home around your smart devices?
   (b) Do you do anything to your devices – such as muting them – to mitigate your security or privacy concerns?
2. What kind of policies or controls would you like to have in your smart home to alleviate your security and/or privacy concerns?

Multi-User Scenarios:

1. How many people live in your home?
2. Who has access to the smart home?
3. Have you ever had disagreements with people in your home about how your smart home is set up?
4. Does everyone who has access have the same level of access? *(If yes:)* Have you had situations where you wanted someone to have limited access, and if so, how did you handle that?
5. Have you ever have situations where houseguests have interacted with your smart home? Did anything go wrong? Did anyone voice any opinions or concerns?

Failures:
Are there any other things that have gone wrong while setting up or using your smart home devices that you'd like to share?

Self-Reporting Technical Skills:
On a scale of 1-5:

1. How familiar are you with technology in general?
2. How familiar are you with computer security?
3. How familiar are you with smart home technology?

Closing Questions:

1. Are there any questions you expected me to ask?
2. Is there anything else you want to tell me about your smart home?

Security Developer Studies with GitHub Users: Exploring a Convenience Sample

Yasemin Acar,*† Christian Stransky,† Dominik Wermke,*† Michelle L. Mazurek,‡ and Sascha Fahl*†
*Leibniz University Hannover; †CISPA, Saarland University; ‡University of Maryland
{acar,wermke,fahl}@sec.uni-hannover.de; stransky@cs.uni-saarland.de; mmazurek@umd.edu

ABSTRACT

The usable security community is increasingly considering how to improve security decision-making not only for end users, but also for information technology professionals, including system administrators and software developers. Recruiting these professionals for user studies can prove challenging, as, relative to end users more generally, they are limited in numbers, geographically concentrated, and accustomed to higher compensation. One potential approach is to recruit active GitHub users, who are (in some ways) conveniently available for online studies. However, it is not well understood how GitHub users perform when working on security-related tasks. As a first step in addressing this question, we conducted an experiment in which we recruited 307 active GitHub users to each complete the same security-relevant programming tasks. We compared the results in terms of functional correctness as well as security, finding differences in performance for both security and functionality related to the participant's self-reported years of experience, but no statistically significant differences related to the participant's self-reported status as a student, status as a professional developer, or security background. These results provide initial evidence for how to think about validity when recruiting convenience samples as substitutes for professional developers in security developer studies.

1. INTRODUCTION

The usable security community is increasingly considering how to improve security decision-making not only for end users, but for information technology professionals, including system administrators and software developers [1, 2, 9, 10, 39]. By focusing on the needs and practices of these communities, we can develop guidelines and tools and even redesign ecosystems to promote secure outcomes in practice, even when administrators or developers are not security experts and must balance competing priorities.

One common approach in usable security and privacy research is to conduct an experiment, which can allow researchers to investigate causal relationships (e.g., [5,8,13,36]). Other non-field-study mechanisms, such as surveys and interview studies, are also common. For research concerned with the general population of end users, recruitment for these studies can be fairly straightforward, via online recruitment platforms such as Amazon Mechanical Turk or via local methods such as posting flyers and advertising on email lists or classified-ad services like Craigslist. These approaches generally yield acceptable sample sizes at an affordable cost.

Recruiting processes for security developer studies, however, are less well established. For in-lab studies, professional developers may be hard to contact (relative to the general public), may not be locally available outside of tech-hub regions, may have demanding schedules, or may be unwilling to participate when research compensation is considerably lower than their typical hourly rate. For these reasons, studies involving developers tend to have small samples and/or to rely heavily on university computer-science students [2, 3, 15, 34, 35, 39]. To our knowledge, very few researchers have attempted large-scale online security developer studies [1, 3].

To date, however, it is not well understood how these different recruitment approaches affect research outcomes in usable security and privacy studies. The empirical software engineering community has a long tradition of conducting experiments with students instead of professional developers [29] and has found that under certain circumstances, such as similar level of expertise in the task at hand, students can be acceptable substitutes [27]. These studies, however, do not consider a security and privacy context; we argue that this matters, because security and privacy tasks differ from general programming tasks in several potentially important ways. First, because security and privacy are generally secondary tasks, it can be dangerous to assume they exhibit similar characteristics as general programming tasks. For example, relative to many general programming tasks, it can be especially difficult for a developer to directly test that security is working. (For example, how does one observe that a message is correctly encrypted?) Second, a portion of professional developers are self-taught, so their exposure to security and privacy education may differ importantly from university students' [32].

The question of how to recruit for security studies of developers in order to maximize validity is complex but important. In this study, we take a first step toward answering it: We report on a experiment (n=307) comparing GitHub contributors completing the same security-relevant

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

tasks. For this experiment, we take as a case study the approach (which we used in prior work [1]) of recruiting active developers from GitHub for an online study. All participants completed three Python-programming tasks spanning four security-relevant concepts, which were manually scored for functionality and security. We found that participants across all programming experience levels were similarly inexperienced in security, and that professional developers reported more programming experience than university students. Being a professional did not increase a participant's likelihood of writing functional or secure code statistically significantly. Similarly, self-reported security background had no statistical effect on the results. Python experience was the only factor that significantly increased the likelihood of writing both functional and secure code. Further work is needed to understand how participants from GitHub compare to those recruited more traditionally (e.g., students recruited using flyers and campus e-mail lists, or developers recruited using meetup websites or researchers' corporate contacts). Nonetheless, our findings provide preliminary evidence that at least in this context, similarly experienced university students can be a valid option for studying professionals developers' security behaviors.

2. RELATED WORK

We discuss related work in two key areas: user studies with software developers and IT professionals focusing on security-relevant topics, and user studies with software developers and IT professionals that do not focus on security but do discuss the impact of participants' level of professionalism on the study's validity.

Studies with Security Focus. In [2] we present a laboratory study on the impact of information sources such as online blogs, search engines, official API documentation and StackOverflow on code security. We recruited both computer science students (40) and professional Android developers (14). We found that software development experience had no impact on code security, but previous participation in security classes had a significant impact. That study briefly compares students to professionals, finding that professionals were more likely to produce functional code but no more likely to produce secure code; however, that work does not deeply interrogate differences between the populations and the resulting implications for validity. In [1], we conducted an online experiment with GitHub users to compare the usability of cryptographic APIs; that work does not distinguish different groups of GitHub users.

Many studies with a security focus rely primarily on students. Yakdan et al. [39] conducted a user study to measure the quality of decompilers for malware analysis. Participants included 22 computer-science students who had completed an online bootcamp as well as 9 professional malware analysts. Scandariato et al. [28] conduct a controlled experiment with 9 graduate students, all of whom had taken a security class, to investigate whether static code analysis or penetration testing was more successful for finding security vulnerabilities in code. Layman et al. [22] conducted a controlled experiment with 18 computer-science students to explore what factors are used by developers to decide whether or not to address a fault when notified by an automated fault detection tool. Jain and Lindqvist [15] conducted a laboratory study with 25 computer-science students (5 graduate; 20 undergraduate) to investigate a new, more privacy-friendly location API for Android application developers and found that, when given the choice, developers prefer the more privacy-preserving API. Barik et al. [4] conducted an eye-tracking study with undergraduate and graduate university students to investigate whether developers read and understand compiler warning messages in integrated development environments.

Studies that use professional developers are frequently qualitative in nature, and as such can effectively make use of relatively small sample sizes. Johnson et al. [17] conducted interviews with 20 real developers to investigate why software developers do not use static analysis tools to find bugs in software, while Xie et al. [38] conducted 15 semi-structured interviews with professional software developers to understand their perceptions and behaviors related to software security. Thomas et al. [34] conducted a laboratory study with 28 computer-science students to investigate interactive code annotations for access control vulnerabilities. As follow up, Thomas et al. [35] conducted an interview and observation-based study with professional software developers using snowball sampling. They were able to recruit 13 participants, paying each a $25 gift card, to examine how well developers understand the researchers' static code analysis tool ASIDE. Johnson et al. [16] describe a qualitative study with 26 participants including undergraduate and graduate students as well as professional developers. Smith et al. [31] conducted an exploratory study with five students and five professional software developers to study the questions developers encounter when using static analysis tools.To investigate why developers make cryptography mistakes, Nadi et al. [25]surveyed 11 Stack Overflow posters who had asked relevant questions. A follow-up survey recruited 37 Java developers via snowball sampling, social media, and email addresses drawn from GitHub commits. This work does not address demographic differences, nor even whether participants were professional software developers, students, or something else.

A few online studies of developers have reached larger samples, but generally for short surveys rather than experimental tasks. Balebako et al. [3] studied the privacy and security behaviors of smartphone application developers; they conducted 13 interviews with application developers and an online survey with 228 application developers. They compensated the interviewees with $20 each, and the online survey participants with a $5 Amazon gift card. Witschey et al. [37] survey hundreds of developers from multiple companies (snowball sampling) and from mailing lists to learn their reasons for or against the use of security tools.

Overall, these studies suggest that reaching large numbers of professional developers can be challenging. As such, understanding the sample properties of participants who are more readily available (students, online samples, convenience samples) is an aspect of contextualizing the valuable results of these studies. In this paper, we take a first step in this direction by examining in detail an online sample from GitHub.

Studies without Security Focus. In the field of Empirical Software Engineering, the question whether or not students can be used as substitutes for developers when experimenting is of strong interest. Salman et al. [27] compared students and developers for several (non-security-related) tasks, and found that the code they write can be compared if they are equally inexperienced in the subject they are working on. When professionals are more experienced than students, their code is better across several metrics. Hoest et al. [14] compare students and developers across assessment (not coding) tasks and find that under certain conditions, e.g., that students be in the final stretches of a Master's program, students can be used as substitutes for developers. Carver et al. [7] give instructions on how to design studies that use students as coding subjects. McMeekin et al. [23] find that different experience levels between students and professionals have a strong influence on their abilities to find flaws in code. Sjoeberg et al. [29] systematically analyze a decade's worth of studies performed in Empirical Software Engineering, finding that eighty-seven percent of all subjects were students and nine percent were professionals. They question the relevance for industry of results obtained in studies based exclusively on student recruits. Smith et al. [30] perform post-hoc analysis on previously conducted surveys with developers to identify several factors software researchers can use to increase participation rates in developer studies. Murphy-Hill et al. [24] enumerate dimensions which software engineering researchers can use to generalize their findings.

3. METHODS

We designed an online, between-subjects study to compare how effectively developers could quickly write correct, secure code using Python. We recruited participants, all with Python experience, who had published source code at GitHub.

Participants were assigned to complete a set of three short programming tasks using Python: an encryption task, a task to store login credentials in an SQLite database, and a task to write a routine for a URL shortener service. Each participant was assigned the tasks in a random order (no task depended on completing a prior task). We selected these tasks to provide a range of security-relevant operations while keeping participants' workloads manageable.

After finishing the tasks, participants completed an exit survey about the code they wrote during the study, as well as their educational background and programming experience. Two researchers coded participants' submitted code for functional correctness and security.

All study procedures were approved by the Ethics Review Board of Saarland University, the Institutional Review Board of the University of Maryland and the NIST Human Subjects Protection Office.

3.1 Language selection

We elected to use Python as the programming language for our experiment, as it is widely used across many communities and offers support for all kinds of security-related APIs, including cryptography. As a bonus, Python is easy to read and write, is widely used among both beginners and experienced programmers, and is regularly taught in universities. Python is the third most popular programming language on GitHub, trailing JavaScript and Java [12]. Therefore, we reasoned that we would be able to recruit sufficient professional Python developers and computer science students for our study.

3.2 Recruitment

As a first step to understanding security-study behavior of GitHub committers, we recruited broadly from GitHub, the popular source-code management service. To do this, we extracted all Python projects from the GitHub Archive database [11] between GitHub's launch in April 2008 and December 2016, yielding 798,839 projects in total. We randomly sampled 100,000 of these repositories and cloned them. Using this random sample, we extracted email addresses of 80,000 randomly chosen Python committers. These committers served as a source pool for our recruitment.

We emailed these GitHub users in batches, asking them to participate in a study exploring how developers use Python. We did not mention security or privacy in the recruitment message. We mentioned that we would not be able to compensate them, but the email offered a link to learn more about the study and a link to remove the email address from any further communication about our research. Each contacted GitHub users was assigned a unique pseudonymous identifier (ID) to allow us to correlate their study participation to their GitHub statistics separately from their email address.

Recipients who clicked the link to participate in the study were directed to a landing page containing a consent form. After affirming that they were over 18, consented to the study, and were comfortable with participating in the study in English, they were introduced to the study framing. We did not restrict participation to those with security expertise because we were interested in the behavior of non-security-experts encountering security as a portion of their task.

To explore the characteristics of this sample, the exit questionnaire included questions about whether they were currently enrolled in an undergraduate or graduate university program and whether they were working in a job that mainly involved Python programming. We also asked about years of experience writing Python code, as well as whether the participant had a background in computer security.

3.3 Experimental infrastructure

For this study, we used an experimental infrastructure we developed, which is described in detail in our previous work [1, 33].

We designed the experimental infrastructure with certain important features in mind:

- A controlled study environment that would be the same across all participants, including having pre-installed all needed libraries.
- The ability to capture all code typed by our participants, capture all program runs and attendant error messages, measure time spent working on tasks, and recognize whether or not code was copied and pasted.
- Allowing participants to skip tasks and continue on to the remaining tasks, while providing information on why they decided to skip the task.

To achieve these goals, the infrastructure uses Jupyter Notebooks (version 4.2.1) [19], which allow our participants to write, run, and debug their code in the browser, without having to download or upload anything. The code runs on our server, using our standardized Python environment (Python 2.7.11). This setup also allows us to frequently snapshot participants' progress and capture copy-paste events. To prevent interference between participants, each participant was assigned to a separate virtual machine running on Amazon's EC2 service. Figure 1 shows an example Notebook.

We pre-installed many popular Python libraries for accessing an SQLite database, dealing with string manipulation, storing user credentials, and cryptography. Table 9 in Appendix C lists all libraries we provided. We tried to include as many relevant libraries as possible, so that every participant could work on the tasks using their favorite libraries.

The tasks were shown one at a time, with a progress indicator showing how many tasks remained. For each task, participants were given buttons to "Run and test" their code, and to move on using "Solved, next task" or "Not solved, but next task." (A "Get unstuck" button was also provided in case the participant accidentally sent Python into an infinite loop or otherwise crashed the Python interpreter running in the Notebook.) After completing (or skipping) all tasks, the participant was redirected to the exit survey.

3.4 Exit survey

Once all tasks had been completed or abandoned, the participants were directed to a short exit survey (cf. Appendix A). We asked for their opinions about the tasks they had completed: Did they think they had solved them? How did they perceive the tasks' difficulty? Did they think their solution was secure? We also were interested in whether they thought about security or privacy when working on the tasks. Finally, we wanted to know whether our participants had worked on similar programming problems in the past. For these task-specific questions, we used our infrastructure to display the participant's code for the corresponding task for their reference. We also asked several questions about demographic information and programming experience, to allow us to distinguish categories of participants.

3.5 Task design

We designed tasks that 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. Most importantly, we designed the tasks to model real-world security and privacy problems that Python developers could reasonably be expected to encounter. While these tasks of course do not represent all possible security tasks, we think they provide an interesting variety for analysis.

URL Shortener

We asked our participants to write code for a method that could be part of a URL shortening service such as bit.ly [1] or Google's URL shortener goo.gl. [2]

[1] cf. `https://bitly.com/`
[2] cf. `https://goo.gl/`

URL Shortener Task

Description:
You are asked to develop code for a URL shortening service similar to `https://goo.gl/`. Users of this service will provide you URLs such as `https://en.wikipedia.org/wiki/History_of_the_Internet`. As a result your service should return a shortened URL such as `http://foo.bar/1xf2`. In this task we would like you to implement a method `shortenURL` that is called for every input URL. The output of this method is a shortened URL for the input URL.

When is the problem solved?
A shortened URL is returned. Please print the output of the method to the console.

While this task does not directly have security implications, we were mainly interested in whether, e.g., they used well established approaches such as message digest or random number generators to generate a short URL, or whether they invented their own algorithm.

Credential Storage

We asked our participants to store login credentials, namely a username and password, in a database backend.

Credential Storage Task

Description:
You are asked to develop a web-application backend that stores login credentials (i.e., usernames and passwords) for the web application's users. In this task, we would like you to implement a method `storeCredentials` that is called for every user at account registration. New login credentials are appended to a local SQLite database. Assume that the username and password are given as HTTP POST parameters to your method. Although we are not asking you to implement the `verifyCredentials` method for authenticating users at this time, assume that you will also be writing that method, so you can choose the storage format within the database. We have prepared a SQLite database named "db.sqlite" containing a table "users" and five text columns, "column1", "column2", "column3", "column4", "column5". You can use any or all of these columns as needed to store users' login credentials; you do not have to use all columns to solve the task.

When is the problem solved?
The credentials are stored in the database file.

This task has direct security implications: we were mainly interested in whether participants followed security best practices. Best practices to store user credentials in a database include hashing and salting the password instead of storing it in plaintext and using some sort of input sanitization (e.g. parameterized instead of raw SQL queries) for SQL queries to prevent SQL injection attacks.

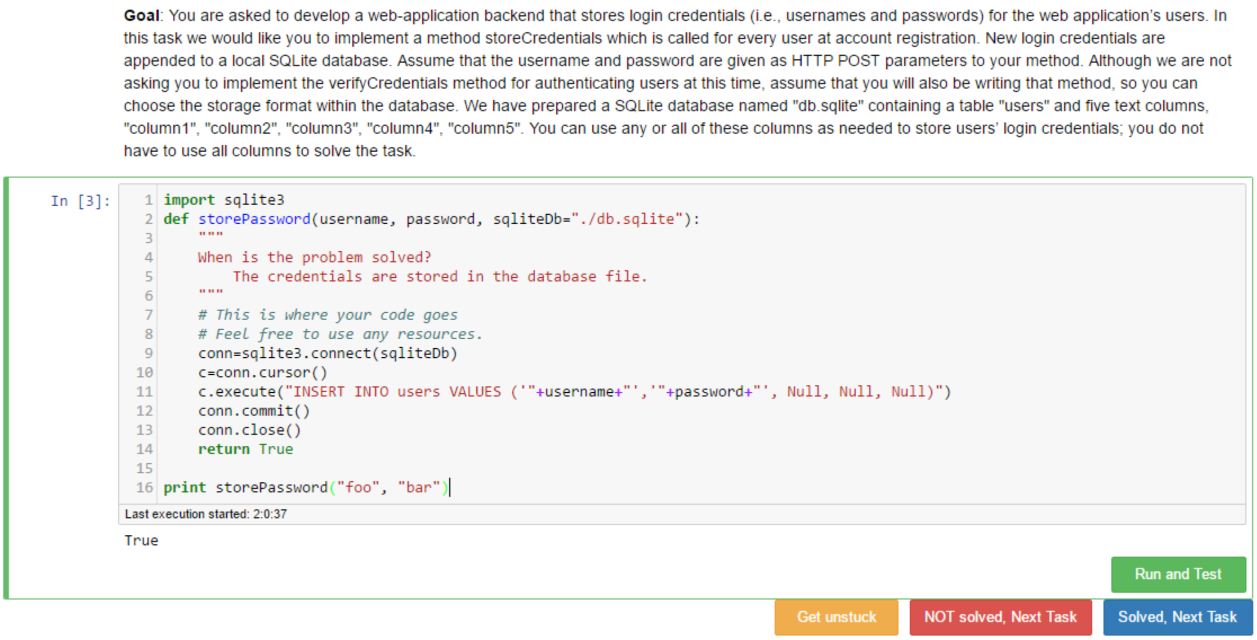


Figure 1: An example of the study's task interface.

String Encryption
We asked participants to write code to encrypt and decrypt a string.

String Encryption Task

Description:
You are asked to write code that is able to encrypt and decrypt a string.

When is the problem solved?
The input string is encrypted and decrypted afterwards. You should see the encrypted and decrypted string in the console.

In this task we were mainly interested in whether participants wrote secure cryptographic code, e.g., choosing secure algorithms, strong key sizes, and secure modes of operation.

For each task, we provided stub code and some comments with instructions about how to work on the task. The code stubs were intended to make the programming task as clear as possible and to ensure that we would later easily be able to run automated unit tests to examine functionality. The code stubs also helped to orient participants to the tasks.

We told participants that "you are welcome to use any resources you normally would" (such as documentation or programming websites) to work on the tasks. We asked participants to note any such resources as comments to the code, for our reference, prompting them to do so when we detected that they had pasted text and/or code into the Notebook.

3.6 Evaluating participant solutions

We used the code submitted by our participants for each task, henceforth called a *solution*, as the basis for our analysis. We evaluated each participant's solution to each task for both functional correctness and security. Every task was independently reviewed by two coders, using a content analysis approach [21] with a codebook based on our knowledge of the tasks and best practices. Differences between the two coders were resolved by discussion. We briefly describe the codebook below.

Functionality. For each programming task, we assigned a participant a functionality score of 1 if the code ran without errors, passed the unit tests and completed the assigned task, or 0 if not.

Security. We assigned security scores only to those solutions which were graded as functional. To determine a security score, we considered several different security parameters. A participant's solution was marked secure (1) only if their solution was acceptable for every parameter; an error in any parameter resulted in a security score of 0.

URL Shortener
For the URL shortening task, we checked how participants generated a short URL for a given long URL. We were mainly interested in whether participants relied on well-established mechanisms such as message digest algorithms (e.g. the SHA1 or SHA2 family) or random number generators, or if they implemented their own algorithms. The idea behind this evaluation criterion is the general recommendation to rely on well-established solutions instead of reinventing the wheel. While adhering to this best practice is advisable in software development in general, it is particularly crucial for writing security- or privacy-relevant code (e.g., use established implementations of cryptographic algorithms instead of re-implementing them from scratch). We also considered the reversibility of the short URL as a security parameter (reversible was considered insecure). We did not incorporate whether solutions were likely to produce collisions (i. e. produce the same short URL for different in-

put URLs) or the space of the URL-shortening algorithm (i. e. how many long URLs the solution could deal with) as security parameters: we felt that given the limited time frame, asking for an optimal solution here was asking too much.

Credential Storage

For the credential storage task, we split the security score in two. One score (password storage) considered how participants stored users' passwords. Here, we were mainly interested whether our participants followed security best practices for storing passwords. Hence, we scored the plain text storage of a password as insecure. Additionally, applying a simple hash algorithm such as MD5, SHA1 or SHA2 was considered insecure, since those solutions are vulnerable to rainbow table attacks. Secure solutions were expected to use a salt in combination with a hash function; however, the salt needed to be random (but not necessarily secret) for each password to withstand rainbow table attacks. Therefore, using the same salt for every password was considered insecure. We also considered the correct use of HMACs [20] and PBKDF [18] as secure.

The second security score (SQL input) considered how participants interacted with the SQLite database we provided. For this evaluation, we were mainly interested whether the code was vulnerable to SQL injection attacks. We scored code that used raw SQL queries without further input sanitization as insecure, while we considered using prepared statements secure.[3]

String Encryption

For string encryption, we checked the selected algorithm, key size and proper source of randomness for the key material, initialization vector and, if applicable, mode of operation. For symmetric encryption, we considered ARC2, ARC4, Blowfish, (3)DES and XOR as insecure and AES as secure. We considered ECB as an insecure mode of operation and scored Cipher Block Chaining (CBC), Counter Mode (CTR) and Cipher Feedback (CFB) as secure. For symmetric key size, we considered 128 and 256 bits as secure, while 64 or 32 bits were considered insecure. Static, zero or empty initialization vectors were considered insecure. For asymmetric encryption, we considered the use of OAEP/PKCS1 for padding as secure. For asymmetric encryption using RSA, we scored keys larger than or equal to 2048 bits as secure.

3.7 Limitations

As with any user study, our results should be interpreted within the context of our limitations.

Choosing an online rather than an in-person laboratory study allowed us less control over the study environment and the participants' behavior. However, it allowed us to recruit a diverse set of developers we would not have been able to obtain for an in-person study.

Recruiting using conventional recruitment strategies, such as posts at university campuses, on Craigslist, in software development forums or in particular companies would likely have limited the number and variety of our participants. As a result, we limited ourselves to active GitHub users. We believe that this resulted in a reasonably diverse sample, but of course GitHub users are not necessarily representative of developers more broadly, and in particular students and professionals who are active on GitHub may not be representative of students and professionals overall. The small response rate compared to the large number of developers invited also suggests a strong opt-in bias. Comparing the set of invited GitHub users to the valid participants suggests that more active GitHub users were more likely to participate, potentially widening this gap. As a result, our results may not generalize beyond the GitHub sample. However, all the above limitations apply equally across different properties of our participants, suggesting that comparisons between the groups are valid.

Because we could not rely on a general recruitment service such as Amazon's Mechanical Turk, managing online payment to developers would have been very challenging; further, we would not have been able to pay at an hourly rate commensurate with typical developer salaries. As a result, we did not offer our participants compensation, instead asking them to generously donate their time for our research.

We took great care to email each potential participant only once, to provide an option for an email address to opt out of receiving any future communication from us, and to respond promptly to comments, questions, or complaints from potential participants. Nonetheless, we did receive a small number of complaints from people who were upset about receiving unsolicited email.[4]

Some participants may not provide full effort or many answer haphazardly; this is a particular risk of all online studies. Because we did not offer any compensation, we expect that few participants would be motivated to attempt to "cheat" the study rather than simply dropping out if they were uninterested or did not have time to participate fully. We screened all results and attempted to remove any obviously low-quality results (e.g., those where the participant wrote negative comments in lieu of real code) before analysis, but cannot discriminate with perfect accuracy. Further, our infrastructure based on Jupyter Notebooks allowed us to control, to an extent, the environment used by participants; however, some participants might have performed better had we allowed them to use the tools and environments they typically prefer. However, these limitations are also expected to apply across all participants.

4. STUDY RESULTS

We were primarily interested in comparing the performances of different categories of participants in terms of functional and secure solutions. Overall, we found that students and professionals report differences in experience (as might be expected), but we did not find significant differences between them in terms of solving our tasks functionally or securely.

4.1 Statistical Testing

In the following subsections, we analyze our results using regression models as well as non-parametric statistical testing. For non-regression tests, we primarily use the Mann-Whitney-U test (MWU) to compare two groups with nu-

[3]While participants could have manually sanitized their SQL queries, we did not find a single solution that did that.

[4]Overall, we received 13 complaints.

meric outcomes, and X^2 tests of independence to compare categorical outcomes. When expected values per field are too small, we use Fisher's exact test instead of X^2.

Here, we explain the regression models in more detail. The results we are interested in have binary outcomes; therefore, we use *logistic* regression models to analyze those results. the consideration whether an insecure task counts as *dangerous*, i.e. whether it is functional, insecure and the programmer thinks it is secure, is also binary and therefore analyzed analogously. As we consider results on a per-task basis, we use a mixed model with a random intercept; this accounts for multiple measures per participant. For the regression analyses, we select among a set of candidate models with respect to the Akaike Information Criterion (AIC) [6]. All candidate models include which task is being considered, as well as the random intercept, along with combinations of optional factors including years of Python experience, student and professional status, whether or not the participant reported having a security background, and interaction effects among these various factors. These factors are summarized in Table 1. For all regressions, we selected as final the model with the lowest AIC.

The regression outcomes are reported in tables; each row measures change in the dependent variable (functionality, security, or security perception) related to changing from the *baseline* value for a given factor to a different value for the same factor (e.g., changing from the encryption task to the URL shortening task). The regressions output odds ratios (O.R.) that report on change in likelihood of the targeted outcome. By construction, O.R.=1 for baseline values. For example, Table 2 indicates that the URL shortening task was 0.45× as likely to be functional as the baseline string encryption task. In each row, we also report a 95% confidence interval (C.I.) and a p-value; statistical significance is assumed for p≤.05, which we indicate with an asterisk (*). For both regressions, we set the encryption task to be the baseline, as it was used similarly in previous work [1].

4.2 Participants

We sent 23,661 email invitations in total. Of these, 3,890 (16.4%) bounced and another 447 (1.9%) invitees requested to be removed from our list, a request we honored. 16 invitees tried to reach the study but failed due to technical problems in our infrastructure, either because of a large-scale Amazon outage[5] during collection or because our AWS pool was exhausted during times of high demand.

A total of 825 people agreed to our consent form; 93 (11.3%) dropped out without taking any action, we assume because the study seemed too time-consuming. The remaining 732 participants clicked on the begin button after a short introduction; of these, 440 (60.1%) completed at least one task and 360 of those (81.8%) proceeded to the exit survey. A total of 315 participants completed all programming tasks and the exit survey. We excluded eight for providing obviously invalid results. From now on, unless otherwise specified, we report results for the remaining 307 valid participants, who completed all tasks and the exit survey.

We classified these 307 participants into students and professionals according to their self-reported data. If a participant reported that they work at a job that mainly requires writing code, we classified them as a professional. If a participant reported being an undergraduate or graduate student, we classified them as a student. It was possible to be classified as either only a professional, only a student, both, or neither. The 307 valid participants includes 254 total professionals, 25 undergraduates, and 49 graduate students. 53 participants were both students and professionals; 32 participants were neither students nor developers. Due to the small sample size, we treated undergraduates and graduate students as one group for further analysis.

The 307 valid participants reported ages between 18 and 81 years (mean: 31.6; sd: 7.7) [Student: 19-37, mean: 25.3, sd: 5.2 - Professional: 18-54, mean: 32.9, sd: 6.7], and most of them reported being male (296 - Student: 21 - Professional 194). All but one of our participants (306) had been programming in general for more than two years and 277 (Student: 18, Professional: 186) had been programming in Python for more than two years. The majority (288 - Student: 20, Professional: 188) said they had no IT-security background nor had taken any security classes.

We compared students to non-students and professionals to non-professionals for security background and years of Python experience. (We compared them separately because some participants are both students and professionals, or are neither.) In both cases, there was no difference in security background (due to small cell counts, we used Fisher's exact test; both with $p \approx 1$). Professionals had significantly more experience in Python than non-professionals, with a median 7 years of experience compared to 5 (MWU, $W = 5040$, $p = 0.004$). Students reported significantly less experience than non-students, with median 5 years compared to 7 years (MWU, $W = 10963$, $p < 0.001$).

The people we invited represent a random sample of GitHub users — however, our participants are a small, self-selected subset of those. We were able to retrieve metadata for 192 participants; for the remainder, GitHub returned a 404 error, which most likely means that the account was deleted or set to private after the commit we crawled was pushed to GitHub. We compare these 192 participants to the 12117 invited participants for whom we were able to obtain GitHub metadata.

Figure 2 illustrates GitHub statistics for all groups (for more detail, see Table 8 in the Appendix). Our participants are slightly more active than the average GitHub user: They have a median of 3 public gists compared to 2 for invited GitHub committers (MWU, W = 1045300, p = 0.01305); they have a median of 28 public repositories compared to 21 for invited participants (MWU, W = 1001200, p < 0.001); they all follow a median of 3 committers (MWU, W = 1142100, p = 0.66); and they are followed by a similar number of committers (10 for participants, 11 for invited; MWU, W = 1146100, p = 0.73).

4.3 Functionality

We evaluated the functionality of the code our participants wrote while working on the programming tasks. Figure 3 illustrates the distribution of functionally correct solutions between tasks and across professional developers and uni-

[5]Some participants were affected by this Amazon EC2 outage: `https://www.recode.net/2017/3/2/14792636/amazon-aws-internet-outage-cause-human-error-incorrect-command`.

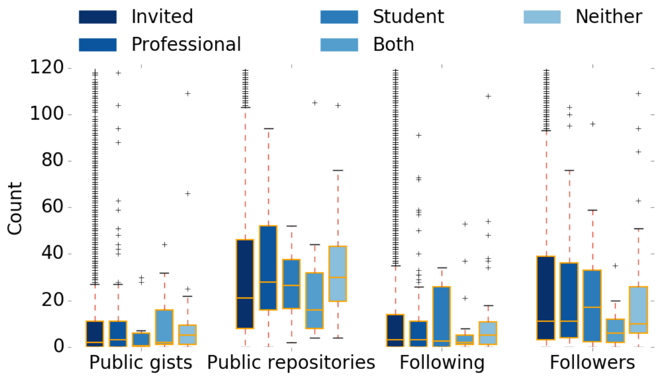


Figure 2: Boxplots comparing our invited participants (a random sample from GitHub) with those who provided valid participation. The center line indicates the median; the boxes indicate the first and third quartiles. The whiskers extend to ±1.5 times the interquartile range. Outliers greater than 150 were truncated for space.

versity students. Overall, professionals got 720 of 804 tasks correct (89.6%), while students got 71 of 84 correct (84.5%); participants who were both students and professionals got 181 of 212 (85.4%) correct, while participants who were neither succeeded in 114 of 128 (89.1%) cases.

Table 2 shows the results of the regression model for functionality. The final model does not include developer or student status, security background, or any interaction effects, suggesting that these factors are not important predictors of functional success. Python experience, on the other hand, did produce a statistically significant effect: each additional year of experience corresponds to on average a 10% increase in likelihood of a correct solution. Comparing tasks, the password storage task proved most difficult: participants were only 0.45× as likely to complete it as to complete the baseline string encryption task. Results for the URL shortening task were comparable to the baseline.

To assess the fit of our regression model, we use Nagelkerke's method [26] to compute a pseudo-R^2 value, somewhat analogous to the standard coefficient of determination commonly used with ordinary linear regression. We find that, relative to a null model that includes only the random (per-participant) effect, our model produces a pseudo-R^2 of 0.07; this is not a particularly strong fit, reflecting the fact that there are potentially many unmeasured covariates, such as the specifics of a participant's prior programming experience and education.

4.4 Security

We evaluated the security of the code based on the codebook described in Section 3.6. In this section, we talk about four tasks instead of three, as the credentials storage task had two security relevant components that we account for individually: secure password digest and SQL input validation (see Section 3.6 for details).

Figure 4 illustrates the distribution of secure solutions between tasks and across professional developers vs. university students. Altogether, professionals got 493 of 720 tasks correct (68.5%), while students got 48 of 71 correct (67.6%); participants who were both students and professionals got 119 of 181 (65.7%) correct, while participants who were neither succeeded in 77 of 114 (67.5%) cases.

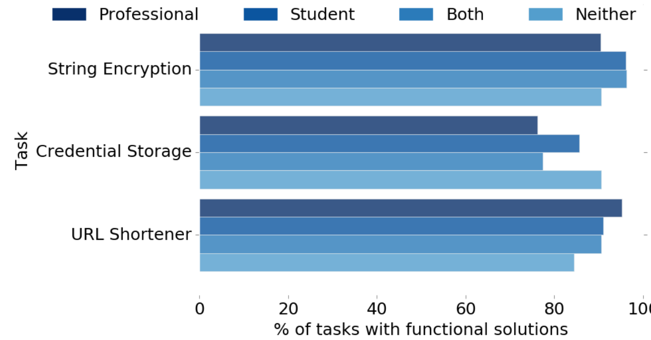


Figure 3: Functionality results per task, split by students vs. professional developers.

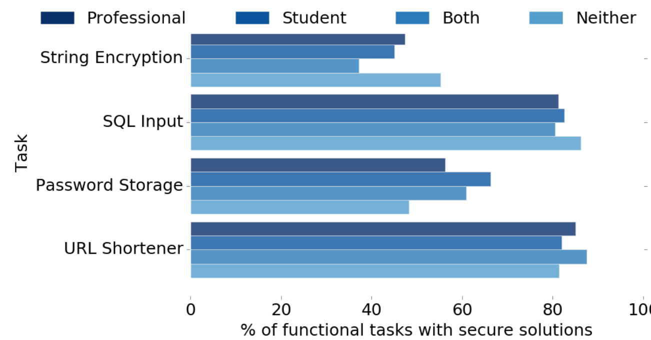


Figure 4: Security results per task, split by students vs. professional developers.

Table 3 lists the results of the final security regression model. This model had Nagelkerke pseudo-R^2 of 0.183, which is a fairly strong fit for an uncontrolled experiment with potential unmeasured factors.

As with the functionality results, none of developer status, student status, security background, nor any interactions, appear in the final model. This again suggests that these factors do not meaningfully predict security success. As before, more Python experience is associated with more success: this time, each year of additional experience adds about 5% to the likelihood of a secure solution. Comparing tasks, string encryption proved significantly more difficult to complete securely than any other task. Password storage was associated with about 2× higher likelihood of success. Both these tasks were significantly harder than SQL input validation and URL shortening. (The non-overlapping confidence intervals indicate significant difference from password storage as well as from the baseline string encryption task). SQL input validation and URL shortening were each about 8× easier to secure than string encryption.

4.4.1 *Security Perception*

We asked participants, for each task, whether they believed their result was secure. In this section, we analyze the incidence of what we call *dangerous* solutions: solutions that

Factor	Description	Baseline
Required		
Task	The performed tasks	String encryption
Participant	Random effect accounting for repeated measures	n/a
Optional		
Python experience	Python programming experience in years, self-reported.	n/a
Security background	True or false, self-reported.	False
Developer	True or false, self-reported.	False
Student	True or false, self-reported.	False
Python experience × task		False:String encryption
Python experience × developer		False:False
Python experience × student		False:False
Developer × task		False:String encryption
Student × task		False:String encryption

Table 1: Factors used in regression models. Categorical factors are individually compared to the baseline. Final models were selected by minimum AIC; candidates were defined using all possible combinations of optional factors, with the required factors included in every candidate.

Factor	O.R.	C.I.	p-value
URL shortener	0.45	[0.22, 0.89]	0.022*
Credentials storage	0.22	[0.11, 0.42]	<0.001*
Python experience	1.10	[1.02, 1.19]	0.014*

Table 2: Results of the final logistic regression model examining functionality of tasks for participants. Odds ratios (O.R.) indicate relative likelihood of succeeding. Statistically significant factors indicated with *. See Table 1 for further details.

Factor	O.R.	C.I.	p-value
URL shortener	8.03	[5.14, 12.53]	<0.001*
Password storage	2.34	[1.6, 3.43]	<0.001*
SQL input	7.69	[4.89, 12.09]	<0.001*
Python experience	1.05	[1.01, 1.1]	0.020*

Table 3: Results of the final logistic regression model examining security of tasks for participants. Odds ratios (O.R.) indicate relative likelihood of succeeding. Statistically significant factors indicated with *. See Table 1 for further details.

are functionally correct and where the participant believes the result is secure, but our analysis indicates that it is not. In a sense, this represents a worst-case scenario, where a developer may confidently release insecure code unwittingly.

Table 4 details how perceptions of security connect to evaluated security. Across tasks, 154 of 1228 (12.5%) solutions were classified as dangerous; happily, dangerous solutions were least common of the four classes, but this rate is still higher than we might hope.

Table 5 reports on a regression model with whether or not a solution is classified as dangerous as the binary outcome. The final model contains no optional factors at all. This indicates that none of Python experience, security background, professional status, or student status is a good predictor of a dangerous outcome. Indeed, the Nagelkerke pseudo-R^2 for this model is only 0.049, which reflects that we did not measure important additional factors.

Our regression model suggests that string encryption, which was most difficult to secure, was (unsurprisingly) also associated with significantly higher likelihood of dangerous solutions than the SQL input and URL shortening tasks. Encryption, however, was comparable to password digests, which also have a cryptographic component. In a prior experiment, we found that about 20% of cryptographic tasks fell into this dangerous category [2].

4.4.2 Investigating Security Errors

We also examined patterns in the types of security errors made by our participants across tasks. Note that these patterns reflect only functional but insecure solutions. In all cases, the same solution may have more than one security error, so percentages generally total to more than 100%.

URL Shortening

First, we consider the URL shortening task. The most common security error (11 cases, 23.0%) was participants who implemented their URL shortening feature using an algorithm that allows an attacker to easily predict the long URL for a given short URL. An example is the use of Base 64 to derive a "short" URL from a given long URL. Although we did not consider keyspace as a security parameter, we briefly review the keyspace generated by participants with functional solutions to this task. 104 participants (37.4%) selected a shortening approach with an unlimited keyspace. The remaining 174 solutions had an average keyspace of 74.1 bits (median 48, standard deviation 6.1). The average for professionals (82.0 bits, median 48) was higher than for students (62.5 bits, median 48), participants who were both students and professionals (58.5 bits, median 36) and participants who were neither (60.6 bits, median 36).

Password Storage

Next, we consider insecure password storage. Here the most common error was hashing the password without using a proper salt, leaving the stored password vulnerable to rainbow-table attacks (74 cases, 77.1%). The second most common error was storing the plain password (45 cases, 46.9%). A total of 19 (19.8%) participants used a static salt

Category	Encryption		Password Storage		URL shortener		SQL input		Total
Dangerous (Perception Secure & Scoring Insecure)	41	(13.4%)	57	(18.6%)	17	(5.5%)	39	(12.7%)	154
Harmless Misperception (Perception Insecure & Scoring Secure)	49	(16.0%)	31	(10.1%)	156	(50.8%)	64	(20.8%)	300
True Positives (Perception Secure & Scoring Secure)	82	(26.7%)	131	(42.7%)	75	(24.4%)	149	(48.5%)	437
True Negatives (Perception Insecure & Scoring Insecure)	135	(44.0%)	88	(28.7%)	59	(19.2%)	55	(17.9%)	337

Table 4: Detailed distribution of perceived and actual security within functional solutions, broken out per task. Percentages are as a function of each task; for example, 13.4% of all encryption solutions were categorized as dangerous.

Factor	O.R.	C.I.	p.value
URL shortener	0.25	[0.12, 0.52]	<0.001*
Password storage	1.16	[0.7, 1.93]	0.565
SQL input	0.53	[0.29, 0.97]	0.038*

Table 5: Results of the final logistic regression model examining perceived security and actual security. Odds ratios (O.R.) indicate relative likelihood of being insecure. Statistically significant factors indicated with *. See Table 1 for further details.

instead of a random salt. Seven (7.3%) participants used MD5, while six (6.3%) used SHA-1 family hashes. Instead of using a one way hash function, four (4.2%) used encryption to secure the password. This is highly discouraged, since an attacker who can gain access to the decryption key is able to recover plain text passwords. These results are detailed in Table 6.

SQL Query

For the SQL query task, 44 (97.8%) of the participants used raw SQL queries instead of prepared statements, leaving their implementation vulnerable to SQL injection attacks. Interestingly, no participant tried to implement their own SQL query sanitization solution.

Encryption

For the string encryption task, one important decision participants made was the choice of cryptographic library (cf. Table 9 for the libraries that came pre-installed). 118 (40.4% of all functional solutions) of the participants used a cryptographic library that was designed with usability in mind, which reduces the necessity to select (and potential make an error with) parameters like algorithm, mode of operation, key size, initialization vector, and padding scheme (cryptography.io: 103, PyNacl: 15, PySodium: 1). 93 participants (31.8% of all functional solutions) chose a more conventional library (PyCrypto: 93), and 73 (25.0% of all functional solutions) used no third-party library at all.

Overall, 15 (12.7%) of participants who applied a usable library made a security error, while 49 (52.7%) of the participants who used a conventional library made a security error. All participants but one who used usable libraries used secure algorithms, modes of operation, and key sizes; the other 14 who made an error used a static initialization vector. Users of conventional cryptographic libraries mostly used a static initialization vector (31 cases, 63.3% of error cases), used an insecure mode of operation (11, 22.4% of error cases), or chose an insecure algorithm (7, 14.3% of error cases). These results indicate that usable libraries do reduce errors, and they are in line with the errors we identified in a prior experiment [1]. These results are detailed in Table 7.

Among participants who did not apply cryptography effectively, 20 used Base64 to encode their plaintext instead of encrypting it, and 43 implemented a very basic substitution cipher like Rot13. An example is shown in Listing 1.

```
def encrypt(plainText):
    return ''.join([chr(ord(c) + n % 5) for
     n, c in enumerate(plainText)])

def decrypt(cipherText):
    return ''.join([chr(ord(c) - n % 5) for
     n, c in enumerate(cipherText)])

stringToEncrypt = "ThisIsAnExample"
encryptedString = encrypt(stringToEncrypt)
print encryptedString
decryptedString = decrypt(encryptedString)
print decryptedString
```

Listing 1: Substitution cipher solution as written by a professional developer participant.

5. DISCUSSION AND CONCLUSIONS

In our online quasi-experiment with 307 GitHub participants, we measured functionality and security outcomes across Python programming tasks. We came into the experiment hypothesizing that whether or not a participant wrote code professionally or as a student would impact at least the functional correctness of their code. However, we found that neither student nor professional status (self-reported) was a significant factor for functionality, security, or security perception. We were also surprised to learn that self-reported security background was equally unimportant. (Note that only small numbers of participants reported that they were exclusively students or that they had a security background, which may affect these results).

We did, however, find a significant effect for Python experience: Each year of experience corresponded to 10% more likelihood of getting a functional result and a 5% better chance of getting a secure result. Differences in experience across students and professionals were significant: Students reported a median of 5 years of experience, compared to 7 for professionals. (On the other hand, experience did not

	Plain password	MD5 hash	SHA1 hash	No salt	Static salt	Raw SQL	Not stored
Professionals	24 (14.0%)	3 (1.7%)	4 (2.3%)	40 (23.3%)	15 (8.7%)	29 (16.9%)	1 (0.6%)
Student	4 (25.0%)	0 (0.0%)	0 (0.0%)	6 (37.5%)	1 (6.2%)	3 (18.8%)	0 (0.0%)
Both	8 (19.5%)	2 (4.9%)	2 (4.9%)	14 (34.1%)	2 (4.9%)	8 (19.5%)	0 (0.0%)
Neither	9 (31.0%)	2 (6.9%)	0 (0.0%)	14 (48.3%)	1 (3.4%)	4 (13.8%)	0 (0.0%)
Total	45 (17.4%)	7 (2.7%)	6 (2.3%)	74 (28.7%)	19 (7.4%)	44 (17.1%)	1 (0.4%)

Table 6: Types of security errors found in functional solutions (and their percentages) by professional, student, both or neither for the password storage task. See Subsection 3.5 for task details and Subsection 3.6 for codebook details.

Library	Used	Weak Algo	Weak Mode	Static IV
Professionals				
No library	44 (22.8%)	42 (21.8%)	0 (0.0%)	0 (0.0%)
cryptography.io	71 (36.8%)	0 (0.0%)	0 (0.0%)	10 (5.2%)
pyCrypto	65 (33.7%)	5 (2.6%)	9 (4.7%)	23 (11.9%)
PyNaCl	10 (5.2%)	0 (0.0%)	0 (0.0%)	1 (0.5%)
Other	3 (1.6%)	3 (1.6%)	0 (0.0%)	0 (0.0%)
Student				
No library	8 (42.1%)	8 (42.1%)	0 (0.0%)	0 (0.0%)
cryptography.io	5 (26.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
pyCrypto	6 (31.6%)	1 (5.3%)	0 (0.0%)	1 (5.3%)
Both				
No library	17 (33.3%)	17 (33.3%)	0 (0.0%)	0 (0.0%)
cryptography.io	16 (31.4%)	0 (0.0%)	0 (0.0%)	3 (5.9%)
pyCrypto	15 (29.4%)	1 (2.0%)	2 (3.9%)	5 (9.8%)
PyNaCl	1 (2.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
pySodium	1 (2.0%)	1 (2.0%)	0 (0.0%)	0 (0.0%)
Other	1 (2.0%)	1 (2.0%)	0 (0.0%)	0 (0.0%)
Neither				
No library	6 (20.7%)	6 (20.7%)	0 (0.0%)	0 (0.0%)
cryptography.io	11 (37.9%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
pyCrypto	7 (24.1%)	0 (0.0%)	0 (0.0%)	2 (6.9%)
PyNaCl	4 (13.8%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Other	1 (3.4%)	1 (3.4%)	0 (0.0%)	0 (0.0%)
Total				
No library	75 (25.7%)	73 (25.0%)	0 (0.0%)	0 (0.0%)
cryptography.io	103 (35.3%)	0 (0.0%)	0 (0.0%)	13 (4.5%)
pyCrypto	93 (31.8%)	7 (2.4%)	11 (3.8%)	31 (10.6%)
PyNaCl	15 (5.1%)	0 (0.0%)	0 (0.0%)	1 (0.3%)
pySodium	1 (0.3%)	1 (0.3%)	0 (0.0%)	0 (0.0%)
Other	5 (1.7%)	5 (1.7%)	0 (0.0%)	0 (0.0%)

Table 7: Types of security errors found in functional solutions (and their percentages) by professional, student, both or neither for the string encryption task. Participant categories are subdivided by the cryptographic library they opted to use. See Subsection 3.5 for task details and Subsection 3.6 for codebook details.

appear to matter for security perception.) This accords well with previous results within the empirical software engineering community (cf. Section 2), which suggest that student and professional developer participants' expertise should be similar to produce similar results. While expertise with Python in our study differs significantly between students and professional developers, their security and privacy expertise is similar (in both cases quite low). At least within GitHub then, it seems that students and professionals can be equally useful for studying usable security and privacy problems, particularly if overall experience is controlled for.

In addition to the small sample size, we speculate that the very similar results across students and professional developers can be accounted for in part because writing security-related code is not something the average software developer deals with on a regular basis, nor is security education a strong focus at many universities teaching computer science. We hypothesize, therefore, that overall these results — experience matters somewhat, but professional status on its own does not — would continue to hold for student and professional populations recruited more traditionally, at local universities and via professional networks. We suspect, however, that typically local university students may have less experience than students recruited from GitHub. Further research is needed to validate these hypotheses.

We found the recruitment strategy of emailing GitHub developers to be convenient in many ways: We were able to recruit many experienced professionals quickly and at a low cost. In addition, many participants expressed to us how much they enjoyed the challenge of our tasks and the opportunity to contribute to our research. However, it does have important drawbacks: we received complaints about unsolicited email from 13 invited GitHub committers and were generally subject to a small opt-in rate. We also found that our participants were slightly more active and therefore not quite representative of the GitHub population; representativeness for professionals (or students) in general is considerably less likely. Overall, the practice of sending unsolicited emails was not ideal, and is unlikely to be sustainable over many future studies. Instead, we plan in the future to develop a GitHub application that would allow developers who are interested in contributing to research to opt in to study recruitment requests, which would benefit both these developers and the research community.

6. ACKNOWLEDGEMENTS

The authors would like to thank Mary Theofanos and the anonymous reviewers for providing feedback; Rob Reeder for shepherding the paper and guiding us in a substantial change of direction; Andrea Dragan and Anne Andrews for help managing multi-institution ethics approvals; Simson Garfinkel and Doowon Kim for contributing to the study infrastructure; and all of our participants for their contributions. This work was supported in part by the German Ministry for Education and Research (BMBF) through funding for the Center for IT-Security, Privacy and Accountability (CISPA), and by the U.S. Department of Commerce, National Institute for Standards and Technology, under Cooperative Agreement 70NANB15H330.

7. REFERENCES

[1] Y. Acar, M. Backes, S. Fahl, S. Garfinkel, D. Kim, M. L. Mazurek, and C. Stransky. Comparing the Usability of Cryptographic APIs. In *Proc. 38th IEEE Symposium on Security and Privacy (SP'17)*. IEEE, 2017.

[2] 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 *Proc. 37th IEEE Symposium on Security and Privacy (SP'16)*. IEEE, 2016.

[3] R. Balebako, A. Marsh, J. Lin, and J. Hong. The Privacy and Security Behaviors of Smartphone App Developers. In *Proc. Workshop on Usable Security (USEC'14)*. The Internet Society, 2014.

[4] T. Barik, J. Smith, K. Lubick, E. Holmes, J. Feng, E. Murphy-Hill, and C. Parnin. Do Developers Read Compiler Error Messages? In *Proc. 39th IEEE International Conference on Software Engineering (ICSE'17)*. IEEE, 2017.

[5] C. Bravo-Lillo, S. Komanduri, L. F. Cranor, R. W. Reeder, M. Sleeper, J. Downs, and S. Schechter. Your Attention Please: Designing Security-decision UIs to Make Genuine Risks Harder to Ignore. In *Proc. 9th Symposium on Usable Privacy and Security (SOUPS'13)*. USENIX Association, 2013.

[6] K. P. Burnham. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, 33(2):261–304, 2004.

[7] J. Carver, L. Jaccheri, S. Morasca, and F. Shull. Issues in using students in empirical studies in software engineering education. In *Proc. 5th International Workshop on Enterprise Networking and Computing in Healthcare Industry (Healthcom'03)*. IEEE, 2003.

[8] S. Fahl, M. Harbach, T. Muders, M. Smith, and U. Sander. Helping Johnny 2.0 to encrypt his Facebook conversations. In *Proc. 8th Symposium on Usable Privacy and Security (SOUPS'12)*. USENIX Association, 2012.

[9] S. Fahl, M. Harbach, H. Perl, M. Koetter, and M. Smith. Rethinking SSL Development in an Appified World. In *Proc. 20th ACM Conference on Computer and Communication Security (CCS'13)*. ACM, 2013.

[10] F. Fischer, K. Böttinger, H. Xiao, C. Stransky, Y. Acar, M. Backes, and S. Fahl. Stack Overflow Considered Harmful? The Impact of Copy&Paste on Android Application Security. In *Proc. 38th IEEE Symposium on Security and Privacy (SP'17)*. IEEE, 2017.

[11] GitHub Archive, Nov. 2 2016. visited.

[12] GitHut: A Small Place to discover languages in github, Nov. 2 2016. visited.

[13] M. Harbach, M. Hettig, S. Weber, and M. Smith. Using Personal Examples to Improve Risk Communication for Security and Privacy Decisions. In *Proc. SIGCHI Conference on Human Factors in Computing Systems (CHI'14)*. ACM, 2014.

[14] M. Höst, B. Regnell, and C. Wohlin. Using Students as Subjects—A Comparative Study of Students and Professionals in Lead-Time Impact Assessment. *Empirical Software Engineering*, 5(3):201–214, 2000.

[15] S. Jain and J. Lindqvist. Should I Protect You? Understanding Developers' Behavior to Privacy-Preserving APIs. In *Proc. Workshop on Usable Security (USEC'14)*. The Internet Society, 2014.

[16] B. Johnson, R. Pandita, J. Smith, D. Ford, S. Elder, E. Murphy-Hill, S. Heckman, and C. Sadowski. A Cross-Tool Communication Study on Program Analysis Tool Notifications. In *Proc. 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE'16)*. ACM, 2016.

[17] B. Johnson, Y. Song, E. Murphy-Hill, and R. Bowdidge. Why don't software developers use static analysis tools to find bugs? In *Proc. 35th IEEE International Conference on Software Engineering (ICSE'13)*. IEEE, 2013.

[18] S. Josefsson. PKCS #5: Password-Based Key Derivation Function 2 (PBKDF2) Test Vectors, Jan. 2011.

[19] Jupyter notebook, Nov. 2 2016. visited.

[20] H. Krawczyk, M. Bellare, and R. Canetti. HMAC: Keyed-Hashing for Message Authentication, Feb. 1997.

[21] K. Krippendorff. *Content Analysis: An Introduction to Its Methodology (2nd ed.)*. SAGE Publications, 2004.

[22] L. Layman, L. Williams, and R. S. Amant. Toward reducing fault fix time: Understanding developer behavior for the design of automated fault detection tools. In *Proc. First International Symposium on Empirical Software Engineering and Measurement (ESEM'07)*. IEEE, 2007.

[23] D. A. McMeekin, B. R. von Konsky, M. Robey, and D. J. Cooper. The significance of participant experience when evaluating software inspection techniques. In *Proc. 20th Australian Conference on Software Engineering (ASWEC'09)*. IEEE, 2009.

[24] E. Murphy-Hill, D. Y. Lee, G. C. Murphy, and J. McGrenere. How Do Users Discover New Tools in Software Development and Beyond? *Computer Supported Cooperative Work (CSCW)*, 24(5):389–422,

2015.
[25] S. Nadi, S. Krüger, M. Mezini, and E. Bodden. "Jumping Through Hoops": Why do Java Developers Struggle With Cryptography APIs? In *Proc. 37th IEEE International Conference on Software Engineering (ICSE'15)*. IEEE, 2016.
[26] N. J. Nagelkerke. A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692, 1991.
[27] I. Salman, A. T. Misirli, and N. Juristo. Are students representatives of professionals in software engineering experiments? In *Proc. 37th IEEE International Conference on Software Engineering (ICSE'15)*. IEEE Press, 2015.
[28] R. Scandariato, J. Walden, and W. Joosen. Static analysis versus penetration testing: A controlled experiment. In *Proc. 24th International Symposium on Software Reliability Engineering (ISSRE)*. IEEE, 2013.
[29] D. I. K. Sjoeberg, J. E. Hannay, O. Hansen, V. B. Kampenes, A. Karahasanovic, N. K. Liborg, and A. C. Rekdal. A survey of controlled experiments in software engineering. *IEEE Transactions on Software Engineering*, 31(9):733–753, 2005.
[30] E. Smith, R. Loftin, E. Murphy-Hill, C. Bird, and T. Zimmermann. Improving developer participation rates in surveys. In *Proc. 6th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE'13)*. IEEE, 2013.
[31] J. Smith, B. Johnson, E. Murphy-Hill, B. Chu, and H. R. Lipford. Questions developers ask while diagnosing potential security vulnerabilities with static analysis. In *Proc. 10th Joint Meeting on Foundations of Software Engineering*. ACM, 2015.
[32] Stack overflow - developer survey results, June 10 2017. visited.
[33] C. Stransky, Y. Acar, D. C. Nguyen, D. Wermke, E. M. Redmiles, D. Kim, M. Backes, S. Garfinkel, M. L. Mazurek, and S. Fahl. Lessons Learned from Using an Online Platform to Conduct Large-Scale, Online Controlled Security Experiments with Software Developers. In *Proc. 10th USENIX Workshop on Cyber Security Experimentation and Test (CSET'17)*. USENIX Association, 2017.
[34] T. Thomas, B. Chu, H. Lipford, J. Smith, and E. Murphy-Hill. A study of interactive code annotation for access control vulnerabilities. In *Proc. 2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC'15)*. IEEE, 2015.
[35] T. W. Thomas, H. Lipford, B. Chu, J. Smith, and E. Murphy-Hill. What Questions Remain? An Examination of How Developers Understand an Interactive Static Analysis Tool. In *Proc. 2nd Workshop on Security Information Workers (WSIW'16)*. USENIX Association, 2016.
[36] B. Ur, P. G. Kelley, S. Komanduri, J. Lee, M. Maass, M. L. Mazurek, T. Passaro, R. Shay, T. Vidas, L. Bauer, N. Christin, and L. F. Cranor. How does your password measure up? The effect of strength meters on password creation. In *Proc. 21st Usenix Security Symposium (SEC'12)*. USENIX Association, 2012.
[37] J. Witschey, O. Zielinska, A. Welk, E. Murphy-Hill, C. Mayhorn, and T. Zimmermann. Quantifying developers' adoption of security tools. In *Proc. 10th Joint Meeting on Foundations of Software Engineering*. ACM, 2015.
[38] J. Xie, H. R. Lipford, and B. Chu. Why do programmers make security errors? In *Proc. 2011 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC'11)*. IEEE, 2011.
[39] K. Yakdan, S. Dechand, E. Gerhards-Padilla, and M. Smith. Helping Johnny to Analyze Malware: A Usability-Optimized Decompiler and Malware Analysis User Study. In *Proc. 37th IEEE Symposium on Security and Privacy (SP'16)*. IEEE, 2016.

APPENDIX

A. EXIT SURVEY QUESTIONS

Task-specific questions: Each task has these questions

On a five-point scale, how much do you agree with the following statements: [strongly agree, agree, neither agree nor disagree, disagree, strongly disagree]

- The task was difficult. (for each task)
- I am confident my solution is correct. (for each task)
- I am confident my solution is secure. (for each task)

What makes this solution either secure or insecure? (free text per task)

When you performed the task, were you thinking about security or privacy? (for each task)

- yes, a lot
- yes, a little
- no

What specifically? (For each task) [free text]

Have you written similar code or come across similar problems in the past? (For each task).

- yes
- sort of
- no

Tell us about it. When was it and what did you do; did you do something differently? [free text]

Demographics and past experience

Check all that apply: Have you ever taken a computer security class?

- at an undergraduate level
- at a graduate level
- via online learning
- via professional training

- another way [specify]
- no, but I took a class that had security as one major component or module
- no

How many computer security classes total have you taken? [input a number]

When did you last take a computer security class? [input a year]

Check all that apply: Do you have experience working in computer security or privacy outside of school?

- Professionally (you got paid to do it)
- As a hobby
- No
- Other [specify]

Check all that apply: Have you ever taken a Python programming class?

- at an undergraduate level
- at a graduate level
- via online learning
- via professional training
- another way [specify]
- no, but I took a class that had Python as one major component or module
- no

How many total Python classes have you taken? [input a number]

When did you last take a Python class [input a year]

Do you have experience programming in Python outside of school?

- Professionally (you got paid to do it)
- As a hobby
- No
- Other [specify]

For how many years have you been programming in Python? [number]

How many Python projects have you worked on in the past? [number]

When did you last work on a Python project? [year]

For how many years have you been programming in general (not just in Python)? [number]

How did you primarily learn to program? (Choose one)

- Self-taught
- In a university / as part of a degree
- In an online learning program
- In a professional certification program
- On the job
- Other [specify]

What is your gender?

- Male
- Female
- Other
- Prefer not to answer

What is your age? [number]

Are you currently a student?

- Undergraduate
- Graduate
- Professional certification program
- Other [specify]
- Not a student

Are you currently employed at a job where programming is a critical part of your job responsibility? [yes/no]

What country did you (primarily) grow up in? [list of countries]

What is your native language (mother tongue)? [list of languages]

B. GITHUB DEMOGRAPHICS

Table 8 compares demographics for invited users vs. participants.

C. INSTALLED PYTHON LIBRARIES

Table 9 lists the Python libraries we pre-installed in the study infrastructure.

	Invited	Valid - Pros	Valid - Students	Valid - Both	Valid - Neither
Hireable	20.5%	19.4%	40.0%	30.6%	23.5%
Company listed	39.4%	43.4%	30.0%	38.9%	17.6%
URL to blog	48.0%	47.3%	40.0%	63.9%	58.8%
Biography added	14.1%	21.7%	20.0%	16.7%	29.4%
Location provided	62.0%	69.8%	50.0%	69.4%	29.4%
GitHub profile creation (days ago, median)	2158	2148	1712	2101	2191
GitHub profile last update (days ago, median)	22	20	23	18	14
Minimal/Maximal age	—	18 / 54	19 / 37	19 / 43	24 / 81
Average age (Std)	—	32.9 (6.7)	25.3 (5.2)	27.5 (4.7)	35.2 (12.7)
More than 2 years programming experience	—	99.5%	100.0%	100.0%	100.0%
More than 2 years Python experience	—	92.5%	85.7%	81.2%	88.7%
Security background	—	6.5%	4.8%	5.7%	6.2%
Male/Female[1]	—	96.5% / 1.5%	100.0% / 0.0%	94.3% / 5.7%	96.9% / 0.0%

[1] the remainder either answered "other" or prefer not to disclose their gender.

Table 8: GitHub demographics for invited users vs. our valid participants.

Library	Version
apsw	3.8.11.1.post1
backports-abc	0.5
backports.shutil-get-terminal-size	1.0.0
bcrypt	2.0.0
blinker	1.3
certifi	2016.9.26
cffi	1.9.1
chardet	2.3.0
configparser	3.5.0
cryptography	1.2.3
cryptography-vectors	1.2.3
decorator	4.0.10
ecdsa	0.13
entrypoints	0.2.2
enum34	1.1.6
file-encryptor	0.2.9
Flask	0.10.1
flufl.password	1.3
functools32	3.2.3.post2
idna	2.0
ipaddress	1.0.16
ipykernel	4.5.2
ipython	5.1.0
ipython-genutils	0.1.0
ipywidgets	5.2.2
itsdangerous	0.24
Jinja2	2.8
jsonschema	2.5.1
jupyter	1.0.0
jupyter-client	4.4.0
jupyter-console	5.0.0
jupyter-core	4.2.0
M2Crypto	0.22.6rc4
m2ext	0.1
macaron	0.3.1
MarkupSafe	0.23
mistune	0.7.3
nbconvert	4.2.0
nbformat	4.1.0

Library	Version
ndg-httpsclient	0.4.0
notebook	4.2.3
passlib	1.6.5
pathlib2	2.1.0
pexpect	4.2.1
pickleshare	0.7.4
prompt-toolkit	1.0.9
ptyprocess	0.5.1
pyasn1	0.1.9
pycparser	2.17
pycrypto	2.6.1
pycryptopp	0.6.0.12...
Pygments	2.1.3
pyinotify	0.9.6
PyNaCl	1.0.1
pyOpenSSL	0.15.1
pysodium	0.6.9.1
pysqlite	2.7.0
python-geohash	0.8.3
python-keyczar	0.715
python-mhash	1.4
pyzmq	16.0.2
qtconsole	4.2.1
requests	2.9.1
simplegeneric	0.8.1
singledispatch	3.4.0.3
six	1.10.0
smbpasswd	1.0.1
ssdeep	3.1
terminado	0.6
tlsh	0.2.0
tornado	4.4.2
traitlets	4.3.1
typing	3.5.3.0
urllib3	1.13.1
wcwidth	0.1.7
Werkzeug	0.10.4
widgetsnbextension	1.2.6
withsqlite	0.1

Table 9: Pre-installed libraries.

Regulators, Mount Up! Analysis of Privacy Policies for Mobile Money Services

Jasmine Bowers, Bradley Reaves, Imani Sherman, Patrick Traynor and Kevin Butler
{jdbowers, reaves, shermani, traynor, butler}@ufl.edu
Florida Institute for Cybersecurity (FICS) Research
University of Florida
Gainesville, Florida

ABSTRACT

Emerging digital financial services use mobile phones to provide access to populations traditionally excluded from the global economy. These "mobile money" services have proven extremely successful in their first ten years of deployment, and provide a powerful means of raising people out of poverty. Such services have access to a wealth of customer information, potentially including entire purchase histories, geolocation, and social network information. In this paper, we perform the first study of privacy policies in mobile money services, evaluating policies from 54 services and comparing them to 50 policies from traditional financial institutions. Because mobile money services are developed under a wide range of regulatory environments, we compare policies to the industry standard (the GSMA's Mobile Privacy Principles) and to a traditional national standard (the FDIC's Privacy Rule Handbook). Our analysis shows that almost half (44%) of these mobile money services do not have *any* privacy policy whatsoever. Of the services that do have privacy policies, roughly one-third (33%) fail to provide them in either of the two most common languages of their market. Furthermore, 50% of these policies do not ever identify to the user what data is actually being collected and stored. Finally, we find that where policies do exist, they are often incomplete and difficult to read by their target customers. These findings show that more work is needed to protect consumer privacy within these mobile money services.

1. INTRODUCTION

Cashless systems underpin our modern economy and the developed world now relies heavily on a massive digital infrastructure capable of moving money across the globe without delay. However, many parts of the world remain unable to easily access these traditional financial networks, often limiting economic expansion and burdening the majority of people around the world with the physical risks and challenges associated with managing currency (e.g., theft, difficulty performing transactions, etc).

"Mobile money" services attempt to address this problem by making phones into payment platforms. Two arbitrary parties, whether in person or at a great distance, can easily transfer money between each other instantaneously. Technologically, this is implemented by various means: built-in "apps" for feature phones, simple text messages exchanged with the mobile money system, and in some limited cases smart phone apps. While conceptually simple, this technology has proven transformative. First, citizens incapable of visiting traditional banking services or maintaining relatively high minimum balances can participate in such services and pay only minimal transaction fees. Second, because virtually anyone with a phone can participate, it is simple for nearly every person and vendor in a country to be enrolled in the service. Finally, many such services are using information gathered on transactions to generate non-traditional creditworthiness measurements and insurance profiles, further enabling those in developing economies to gain access to investments that have proven essential for raising people out of poverty [39].

The implications of collecting and managing customer data in this environment are more risky than in traditional financial services. Specifically, because true peer-to-peer payment is enabled by mobile money services, they learn both their customers' entire financial history and their social network. Moreover, many services also collect supplementary information including geographic location and the names of other applications installed on a device. The need for strong, clearly written privacy policies is therefore evident, and has been strongly supported by the industry group (i.e., the GSMA) since 2011 [43]. However, how the industry has adopted such regulation has not previously been explored.

In this paper, we perform the first independent analysis of privacy policies for the mobile money industry. While work has been done to evaluate privacy policies of traditional financial services in the past [13, 19, 37], our work is different for a number of reasons. First,

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

ours is the first work to evaluate mobile money services, which are generally run by non-traditional financial entities (e.g., telecommunications providers) and are therefore not subject to the same kinds of regulations as the conventional banking industry. Second, we are not attempting to measure the "goodness" of the standards for the industry; rather, our work seeks to measure mobile money services against standards that are already in place. We believe this is critical as it measures the health of consumer privacy within the industry as expectations are currently set. Arguments for greater protections beyond what is suggested can always be made. Finally, mobile money services served approximately 411 million customers in 2015, and recent moves including demonetization in India put the industry's population base in the billions [38]. As these customers often represent many of the world's most financially vulnerable, it is critical for the privacy community to ensure that they are adequately protected.

We make the following contributions:

Compare Policies Against GSMA and FDIC Recommendations: Given that mobile money services are deployed across a wide range of regulatory environments, performance against a single national standard for privacy policies would be incomplete. Accordingly, we determine how these services conform to the mobile money industry (i.e., the GSMA) and widely used government (i.e., the US Federal Deposit Insurance Corporation (FDIC)) recommendations for communicating privacy policies. Our analysis shows that mobile money services lag significantly behind traditional financial institutions, with 44% having no available policy. Moreover, where policies are available, their coverage of critical topics is often limited. All but one service fail to meet either standard.

Measure Readability: Literacy rates are often lower in the developing world. To measure whether or not privacy policies adjust for this difference, we use multiple tests of readability from the linguistics community to compare traditional financial and mobile money offerings. We show that, on average, privacy policies for mobile money services have a higher grade-level readability score than traditional banks (12.1 vs 10.8), meaning that such policies are in general harder to understand.

Measure Availability in Official Languages: Given that mobile money services are available in a wide array of countries, we would expect that we would see privacy policies available in a wide array of languages. Unfortunately, our study shows of the mobile money services that have a privacy policy, 33% of them do not have have a version written in the first or second official language of their country of operation.

We believe that our results point to a critical problem for mobile money services: privacy policies are not widely available, and where they are, the majority of them suffer from low readability and coverage issues. By providing the first such study of these policies, we believe that government and industry regulators can (and must) better address this problem.

The remainder of the paper is organized as follows: Section 2 discusses related research; Section 3 provides background on mobile money; Section 4 describes our methodology; Section 5 provides our results; Section 6 further discusses the implications of these results, and Section 7 provides concluding remarks.

2. RELATED WORK

Mobile money services have had enormous positive impact in enabling financial inclusion in the developing world. However, the security provided by such services has recently come into question. First generation services, which rely on SMS or USSD channels in 2G networks, are inherently insecure due to the use of weak (or no) ciphers on the air interface and the lack of strong end-to-end encryption [10,28]. Unfortunately, Reaves et al. [31] demonstrated systemic problems in smart phone apps throughout the ecosystem, including poor configuration, failure to properly authenticate certificates and in some cases, a complete lack of any protections. Moreover, most developers failed to improve the security of their applications over a year later in spite of receiving detailed vulnerability reports [30]. Castle et al. interviewed developers to understand the cause of such weaknesses, learning that many developers had difficulty properly using security libraries [11].

Creating "good" privacy policies is a challenge in and of itself. Entities handling user data must strike a careful balance between both comprehension and comprehensiveness. A number of researchers [21, 32, 36] provide guidelines for policy creation. McDonald and Cranor [25] argued that a singular focus on coverage represents a significant expense to users and determined that it would take users over 200 hours per year to read the policy of each website visited. Other researchers have attempted to achieve such balance, using techniques such as bulleted lists [17], privacy "nutrition labels" [22], and natural language processing for minimization [46]. Unfortunately, while such techniques appear to improve readability, only one (Federal Reserve privacy notice template) has been adopted at large scale [4]. Our work is able to bypass this challenge to some extent, as it seeks to measure adherence to widely known standards.

While Reaves et al. [31] did evaluate the terms of service for mobile money services they studied, the state of privacy policies in the mobile money environment has not previously been examined. Academic analysis of privacy policies in financial services dates back to at least 2001, with Hochhauser's study demonstrating that understanding the privacy policies of the top 60 US-based financial institutions required an average reading level of 15.6 (i.e., the reading level of a 3rd year college student) [19]. Jensen and Potts [20] measured readability and accessibility in the privacy policies of well-known websites, similarly finding a great need for improved readability. More recently, Cranor et al. [13] conducted a large-scale study of 6,191 US-based financial institutions and focused on their policies for third parties, reasons for data sharing, and opt-out. Closest to our work is that of Sheng et al., [37] which found a lack of significant improvements to privacy policies after the passage of the Gramm-Leach-Bliley (GLBA) Act [45], which mandated that banks make clear how they handle private customer data.

3. MOBILE MONEY

Billions throughout the developing world lack access to even the most basic financial services, and this especially includes many of the world's poor [7]. Financial exclusion result in difficulty receiving wages, government transfers, remittances, making payments, and transferring money to local friends or relatives. This is to say nothing of the lack of simple conveniences provided by the modern financial services. There are a number of reasons why the poor are excluded from traditional financial services, including account fees, difficulty conducting business during relatively limited banking hours, and simple lack of available services (especially in rural areas).

In recognition of this problem, governments and development agencies are embarking upon programs to improve financial inclusion. These efforts are worthwhile because making saving and transferring money easier gives participants the ability to better support themselves as well as provide a safety net for family and friends. In many cases, such safety nets prevent minor financial setbacks from becoming personal crises.

One of the barriers to financial inclusion is that traditional brick-and-mortar banking comes at a high overhead, and it is simply not economical to provide services to customers with low transaction volume or balances. As a result, governments and NGOs are turning to a new model for financial inclusion: digital financial services served through ubiquitous mobile devices.

Services that provide the ability to store value and make payments through mobile phones are often called "mobile money services." The first such service, M-Pesa, was deployed by SafariCom in 2007. M-Pesa pioneered a model where users could send and receive payments directly from their mobile phone, as well as deposit and withdraw funds from an account at any local airtime vendor. This model quickly achieved enormous success, and by 2013 M-Pesa supported payments amounting to a third of Kenyan GDP [26]. Other carriers and third party providers have taken notice, and supported by development organizations like the Gates Foundation and the World Bank, industry consortia like the GSMA, as well as motivated by their own commercial interest, mobile money services have been deployed in developing countries worldwide. Mobile money services have been augmented by other financial services – notably microfinance (small loans) and even life insurance. Figure 10 in Appendix 7 shows the EcoCash mobile app payment interface.

Mobile money services are distinct from both traditional mobile banking (i.e., phone-based access to traditional banking accounts) and many popular mobile payment services in developed countries (e.g., PayPal, Venmo). Figure 1 highlights the most important differentiators.

First and foremost, mobile money is distinct because

	Mobile Money	**Mobile Banking & Mobile Payments**
Target Audience	Unbanked consumers	Existing customers
Financial System	Non-interoperable, closed-loop	Global interoperable financial network
Regulators	Financial, Telecom, or other agencies	Financial regulators only
Technology	SMS, USSD, Mobile Internet	Mobile Internet Only

Figure 1: Mobile money services are distinct from traditional mobile banking and mobile payments in several important ways.

of the market it serves. Mobile money services primarily target unbanked customers new to financial services, while traditional mobile banking and payment services focus on providing new features to existing customers. Second, mobile money is neither based on nor is currently interoperable with the existing traditional banking service. In fact, mobile money services are only rarely interoperable with one another, even within the same market. By contrast, both traditional mobile banking and mobile payment services like Apple Pay interact with the global banking network.

Mobile money services are also technologically unique. Some services are available exclusively through mechanisms compatible with feature phones like SMS or Universal Supplementary Services Data (USSD) in the cellular network. As smartphones increase in popularity in developing countries, more mobile money services are deploying smartphone apps. Many of these services, regardless of whether they operate using legacy channels, smartphone apps, or both, often expose their users to avoidable security vulnerabilities that place their users at risk of loss of funds or disclosure of personal information [11, 31]. This is particularly troubling because the terms of service of most of these services hold users accountable for losses due to fraud, in contrast to regulations and policies like those in the United States, which limit the consumers' exposure to and liability for loss.

In fact, regulatory structures are another area where mobile money is distinctive compared to traditional banking. Many governments are enabling and encouraging mobile money growth by significantly relaxing the regulatory requirements for mobile money services. For example, many "Know Your Customer (KYC)" and anti-money laundering and countering financing of terrorism (AML/CFT) regulations are significantly relaxed to enable simple and practical enrollment of mobile money users. These regulation relaxations are necessary for

Figure 2: Countries represented in our selection of all mobile money services known to have Android apps.

mobile money to be successful, and AML/CFT goals are typically addressed by simply limiting the balances users are allowed to accumulate. However, the need for relaxed regulations in registration does not mean that regulations in other areas – particularly for data security, fraud liability, and data privacy – are unnecessary. Many mobile money services use transaction data to generate non-traditional creditworthiness measurements for users in developing economies.

4. METHODOLOGY

Our research seeks to address three main questions. First, *do mobile money applications have privacy policies and, if so, what do they cover?* Second, *are these policies written such that they are understandable by their target audience?* Finally, *are these policies written in the most commonly spoken languages in the country in which the application is deployed?*

To answer these questions, we began by collecting privacy policies of the top US banks by assets to serve as a reference group. We chose these banks because they have a well-understood regulatory structure. Because a similar set was studied in prior work [19], we argue that our observations in this space serve as a comparison point for emerging digital financial services. We then collect all available privacy policies for all 54 mobile money applications known to have an Android app in 2016. We focus on services with Android apps because smartphone applications have the ability to collect extremely fine-grained data about their users' behavior, in contrast with the limited data collection made possible with feature-phone based services.

We then manually code these policies to investigate compliance with the industry-wide guidance provided by the GSMA as well as the guidance provided by the FDIC. The GSMA Mobile Privacy Principles represent an international accord on privacy polices agreed to by the industry trade group. The GSMA principles have been publicly available since 2011 [43], thereby allowing sufficient time for mobile money services to incorporate its requirements. We note that the GSMA has also released a Code of Conduct specifically for mobile money providers [42]. This document directly addresses issues of user privacy, and explicitly calls on mobile money services to ensure the following principles: "Governance", "Transparency and Notice", "User Choice and Control" and "Minimization of Data Collection and Retention". We decided to use the more complete set of Mobile Privacy Principles [43] for three reasons. First, the Code of Conduct has only been publicly available since 2014, three years fewer than the Mobile Privacy Principles document. Second, with the exception of "Governance", each of the practices in the Code of Conduct maps directly to a principle in the Mobile Privacy Principles. "Governance" most readily maps to "Accountability and Enforcement". Third, because mobile money services comprise mobile applications, the more explicit Mobile Privacy Principles apply to them as well.

Why use the FDIC principles: The FDIC principles [16] provide a more comprehensive standard by which to evaluate privacy statements than the GSMA standards. It was simply not practical to judge each policy based on the standards of the 32 different countries for which we collected mobile money apps. Accordingly, to have an objective basis of comparison, we chose to standardize our analysis on the FDIC standards and the GSMA standards. Including the FDIC principles allowed us to compare the policy coverage of US banks and mobile money applications to both US and international standards, although we understand that there are many confounds that will affect this comparison. We do not claim that the FDIC principles are an ideal standard, but an objective one that has been widely examined. It is also one we believe to be reasonably comprehensive. We readily admit that existing mobile money policies may not have been written with FDIC standards in mind; at the same time, the FDIC principles are general enough that we believe that the content they cover should be in any good privacy policy.

We conclude with an analysis of readability of policies and an analysis of availability of mobile money policies in dominant languages in their respective countries.

4.1 FDIC and GSMA Regulations

The GSMA and FDIC identify key principles that all privacy policies should adhere to and include. Below are the 11 principles used in our privacy policy analysis:

GSMA

Purpose of Data Collection: Policies should disclose the purpose of collecting, accessing, and sharing user data and ensure that each purpose is for legitimate business operations.

Children and Adolescents: If applicable to children, the service should guarantee that the child's personal information is properly collected and should abide by all laws related to children's privacy.

Accountability and Enforcement: Employees are held accountable for proper use and protection of user data.

FDIC

Collection Process: Notices should list the types of personal information that is collected.

Definitions: Notices should terms concerning collection process, information disclosure, etc.

Examples: Notices should include examples of the collection process, information disclosure, etc.

Third Parties: Notices should disclose affiliates that the bank shares nonpublic personal information with.

FDIC & GSMA

User Choice and Control: Notices should disclose the user's right to opt-out and how users can control the use of their personal information.

Security: Notices should disclose how personal information of users will be protected and safeguarded.

Sharing Process: Notices should include the personal information of users that may be disclosed.

Data Minimization/Retention: Information sharing practices of personal information of former customers should be disclosed and only the minimum amount of user information should be collected, accessed, and used at all times.

4.2 Selection and Collection Process

We compiled a list of the top 50 U.S. banks (by assets) based on the Federal Reserves Statistical Release of Large Commercial Banks [5]. We used the GSMA Mobile Money Tracker [18] to identify mobile money services, then manually searched the Google Play market and mobile money provider websites to locate those with smartphone applications. While a large number of applications exist, we carefully inspected each candidate application to ensure that it actually provides payment functionality. We identified 54 such services from 32 countries, as shown in Figure 2.

We located and downloaded privacy policies for all of the banks and mobile money services studied. Our search process was systematic and exhaustive, and is described below.

For a mobile money service, we first visit the app's Google Play Store profile to determine if a privacy policy link exists. If yes, we downloaded the policy from the Play Store. Otherwise we visit the website link in the Play Store (if present) or search Google for the website. If a website was not found after the Google search, the app was marked as not having a policy.

On the website for a mobile money service or bank, we first search for app's privacy policy on the main page. If not found, we searched the website to locate a policy. As a last resort, we examined the "About" pages of the website, then directly searched Google for a policy. Finally, if no policy had been found, we mark the bank or service as not having a policy.

We gathered every privacy notice/policy available on each bank or service's website, including documents termed "privacy policy," "privacy notice," "consumer privacy policy," "cookie policies," "online privacy policies," and "mobile privacy policies". Where applicable, we also investigated "terms of service" documents. We collect and analyze all privacy related documents, because some of the banks used the terms "privacy notice" and "privacy policy" interchangeably.

4.3 Coding Process

We conducted a manual coding analysis to determine which principles our collected policies adhered to.

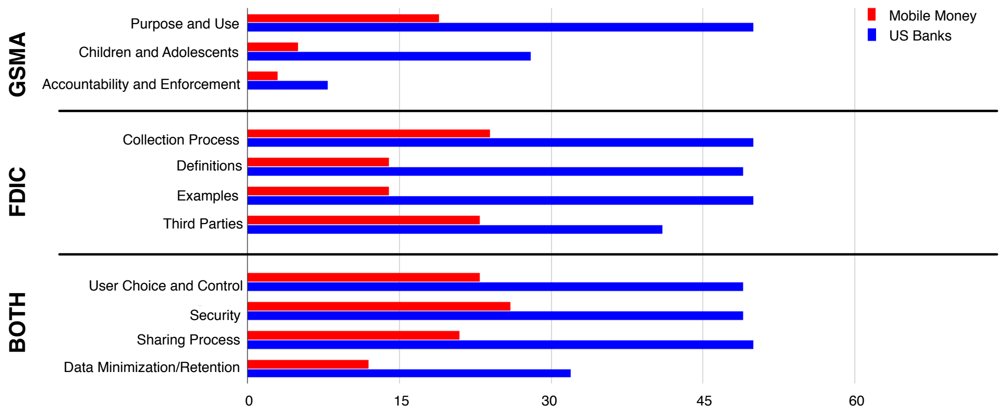
Our coding analysis consisted of two key phases: an initial key word search and subsequent manual analysis. Before coding, we generated a codebook that directly correlated to our policy principles. For example keywords for the **user options** principle included *disable*, *edit*, *user can request*, *user can edit*, *user can change*. We show our codebook in Appendix 7 in Table 3.

We note that a document only needed to *simply mention* the principle to receive credit for covering that principle. We do not otherwise evaluate the extent to which we believe the policy adheres to the letter or spirit of the recommendations. Because our work primarily is concerned with whether policies cover the requisite data *at all,* the absence of a principle in our reports is a strong indicator important privacy issues are being ignored by a particular policy. Thus, the keyword analysis was sufficient to show that most of the mobile money policies failed to mention why data was collected.

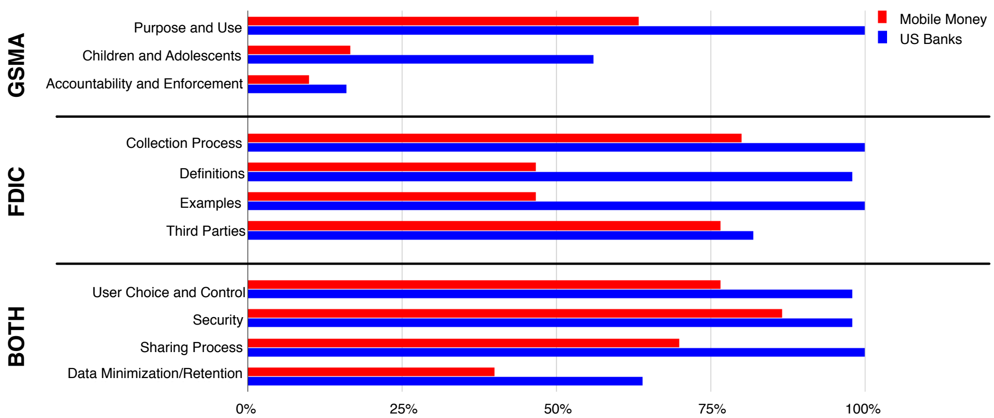
Two of the authors, both PhD students with a prior graduate course in privacy communication, served as the coders. Coders were provided with a digital copy of each policy document. A Google Form was created to streamline the coding process and eliminate error. The coders were instructed to score each policy based on the 11 principles and their corresponding keywords. During the coding process, if the coder did not find exact keywords in the codebook but did find similar text, the coder was instructed to use their best judgment when scoring that principle. The coders were instructed to only assess the policy document and not any other resources (e.g., website.) If neither the keywords nor similar text were found for a specific principle, the policy received a score of zero for that principle.

Once all documents were analyzed individually, we combined the results for every bank or service so that if any of a bank or service's documents discussed a principle, we consider that bank or service as having a documented policy for that principle. For example, Bank of America's Consumer Privacy Notice discusses data minimization/retention, while the Cookie Guide does not. Thus, we mark Bank of America as having a policy for data minimization since at least one of its documents discussed the principle.

The results of the independent trials were compared and mutually reconciled to arrive at the reported data. During the reconciliation process, if the results of the coders differed, we discussed the instructions and thought process with both coders to determine the final score for each policy and principle.



(a) Total Number of Systems



(b) Percentage of Policies Found

Figure 3: Representation of policies that cover each principle in the GSMA, FDIC, or both sets of recommendations. For every principle, banks outperform mobile money services

We computed Cohen's Kappa to measure inter-rater reliability on codings of each policy. Coders agreed "substantially" (κ>0.8) on five categories: Purpose and Use, Children and Adolescents, Definitions, Examples, and Security. In addition, the results of five other categories (Accountability and Enforcement, Collection Process, Third Parties, User Choice and Control, and Data Minimization/Retention) were classified "moderate" (0.41-0.60) and one (Sharing Process) was classified "slight" (0-0.20) [23]. In this latter case, the coders differed in their interpretation of whether the Sharing Process principle was met. One coder gave the policy credit if it simply mentioned sharing, while the other coder looked for a more concrete process (e.g., the sentence "We will limit the access, collection, sharing, disclosure, and further use of your Personal Information to meeting legitimate business purposes or to otherwise meet legal obligations" in the GCash policy was interpreted differently). Coders reconciled their differences by agreeing to adopt the broadest reasonable interpretation of the principle. The same procedure occurred with respect to categories with moderate kappa values.

4.4 Readability Analysis

The linguistics community has created a number of metrics that compute a document's readability score indexed to "grade level" expectations of reading ability.While there are a number of such scores, including the well-known Flesch-Kincaid score, the linguistics community has yet to identify a single, "gold-standard" technique. Accordingly, it has become common (including within the literature on privacy policies [8] and medical documents [9,15,40]), to use more than one readability scoring mechanism in their study.

To measure the readability of our privacy policies, we first manually condense the various policies into individual text-only documents to be analyzed. This manual canonicalization ensured that formatting and typography (e.g., two-column documents) did not prevent an accurate assessment. We then calculated a number of properties of the submitted text, including five readability scores, the estimated reading time, and summary statistics like word counts. We computed the Flesch-Kincaid Grade Level score, the Gunning-Fog score, the Coleman-Liau Index, the SMOG Index, and the Automated Readability Index. Large, distinct documents had their readability assessed independently, then averaged to produce the final score for a bank or service.

4.5 Language Analysis

Many mobile money services operate in nations that are multilingual, and it is important that the policies are provided in languages users actually can read.

Mobile money is a global phenomenon, but practically all services serve a single country. Because language needs vary in different countries, we searched the website of each mobile money service for the availability of privacy policies in multiple languages. To enable an objective comparison among services, we search for only the top two languages in the relevant country as reported by the CIA World Fact Book [6]. In some cases, the World Fact Book reports actual percentages of population that speak a particular language; we use those figures to identify the "top two" languages where available. When unavailable, we use the first two languages listed in the World Fact Book. We also found that three countries with mobile money services (Uruguay, Dominican Republic and Brazil) only had one official language, and our reported results reflect this.

We note that our analysis of policy details is limited to the available English-language policies. A deeper analysis of non-English policies was not possible given that we found documents written in more than 10 languages. Additionally, because automated translation is known to be an open problem, we did not attempt to use such tools to translate such documents as they were likely to unfairly create errors.

5. RESULTS

The individual banks and services studied and their performance on all 11 guidelines are presented in Table 1 and Table 2 in Appendix 7. A summary of the counts of both banks and mobile money services adhering to each principle are are shown in Figure 3a and Figure 3b. In the figures, we present whether mobile money systems adhere to the recommendations we describe in two ways. The first is a total count basis: "of the systems we analyze, how many provide the recommended disclosures to their users." This reflects the ability of mobile money users to learn about how their data is used. The second is on a percentage basis: "of the policies we have, how many adhere to the recommendations for privacy disclosure." This second graph indicates the overall coverage of policies *when they exist.* We provide the first analysis of how privacy is guaranteed in mobile money services and how it compares to privacy in an established banking ecosystem.

5.1 Availability and Freshness

Our first analysis looks at the availability and update frequency of privacy policies for both top US Banks and mobile money services.

Policy Availability: We find that all of the 50 US banks provided at least one privacy policy document. However, only 30 out of the 54 mobile money services we examined had a privacy policy, meaning that 24 services, or 44.4% of all mobile money services with smartphone apps, had *no privacy policy of any sort*, rendering any analysis of their handling of sensitive data impossible. This means that neither privacy experts nor end users have any knowledge of the data practices of these services, much less any rights or guarantees about how that data will be used.

Update Frequency: Privacy policies should be regularly updated to ensure that they still reflect current data handling practices. While the FDIC requires all banks to develop and disseminate updated privacy policies at least once in any 12-month period [16], we found that only 30 of the 50 banks had policies that had been updated within the past year. Mobile money services demonstrated less frequent updates of privacy policies. Of the 30 services that had a privacy policy, only 9 (30%), included information about when they were either written or last updated. Of these 9 services with dated policies, only 5 (17%), had policies that had been updated within the past year.

These first two measures already indicate an important difference between the practices of traditional financial institutions. While both sectors could improve their performance in keeping policies updated, traditional financial institutions far outperform mobile money services in making policies available.

5.2 Policy Content

We next examined the content of the privacy policies we obtained to determine how they adhered to both national regulators (FDIC) and industry guidelines (GSMA).

Substantially more banks adhered to every principle we track than mobile money services. Only one mobile money service, GCash, covers all principles. This is surprising because many more banks conformed to the GSMA recommendations than mobile money services. It is important to note that the GSMA is a consortium of mobile phone carriers, and banks are not members. However, many mobile money services are operated by carriers that *are* members of this organization. This means that US banks have a significantly higher rate of adherence to a standard that they are *not party to* than an industry that has agreed to implement the standard.

(GSMA) Purpose and Use: The *purpose of data collection* information is critical to users. While every bank privacy policy indicated the purpose of data collection, only 19 mobile money services, or 63% of services that have any privacy policy, indicate *why* data is actually being collected. The remaining services give *no indication* as to the purpose of data collection.

(GSMA) Children and Adolescents: The GSMA Principles recommend that any service intended for use by children have special policies for the data collected by child users. We believe that mobile money services operate in a "gray area" in this respect. Mobile money services are not intentionally marketed for children, but where mobile money services are commonly used, children will likely use these services. We note that there is very limited consideration of children's privacy amongst mobile money service policies, and it is only mentioned in 5 mobile money policies, in contrast to 27 bank policies. However, in both cases a substantial number of policies make no mention of children (83.33% of mobile money service policies and 46% of bank policies, respectively). We believe both mobile money services and US banks should consider this issue more seriously.

(GSMA) Accountability and Enforcement: GSMA principles charge employees with a duty to maintain data privacy according to the the privacy policy. However, only 8 US banks and 3 mobile money privacy policies have any mention of the obligations of employees. We note that many policies may not explicitly discuss this principle, presuming employees will be responsible for implementing published policies.

(FDIC) Collection Process: The FDIC recommends that financial services disclose what personal data is collected. While mobile money services tended to at least mention this principle at a higher rate than other principles (24 services, or 80% of available policies), they still fall short of the 96% coverage rate of US banks.

(FDIC) Definitions and Examples: *Definitions* and *examples* are key components of privacy policies because they give the user a clear understanding of the terms used in the policy and the specific information that will be used throughout their interaction with the bank/service. In the most significant quantitative difference seen between the two groups of policies, almost all bank policies provided definitions of the data (47 policies) that would be collected from users and gave examples of its usage (48 policies). In contrast, only 8 mobile money service policies defined what type of data would be collected and only 14 policies gave examples of its usage. Even though policies may technically inform users of the data being collected, they can be obfuscated such that the data collected and shared is difficult to define. *We find the majority of mobile money services with a privacy policy fail to identify to their users what data is actually being collected or stored.*

(FDIC) Third Parties and (Both) Data Sharing Practices: Notices of third party data sharing practices are another critically important aspect of privacy policies. Our measurement of principles distinguishes between whether third party interactions are discussed ("Third Parties") and whether there are additional details about what is being shared with third parties and why ("Sharing Process"). We find that 24 mobile money services (80% of those with policies) address third parties in their policies, and this is actually at a rate comparable to the 41 US banks (82%). However, when it comes to details of the sharing process, there is a significant disparity between our two populations. All US banks discuss this issue, compared to only 21 mobile money services (70% of available policies). This disparity is troubling, as sharing of data with unaffiliated third parties is a serious concern for customers.

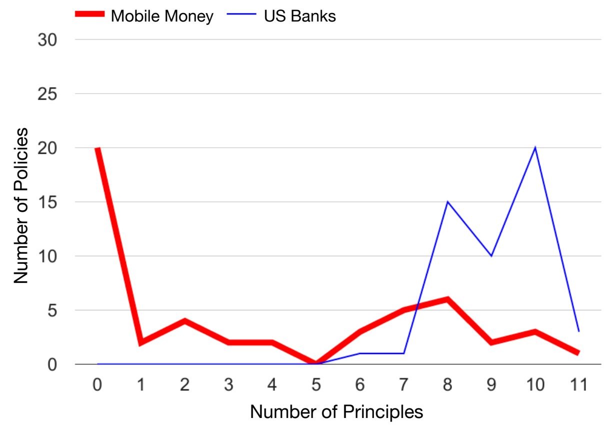


Figure 4: The total number of principles covered by banks and mobile money services. While the US banks meet most of the FDIC and GSMA principles, the mobile money industry falls far short of these standards.

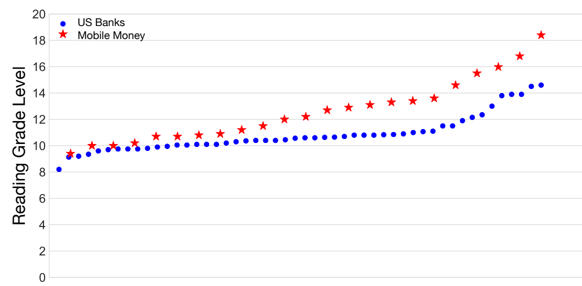


Figure 5: Reading Grade Level scores of U.S. banks and Mobile Money services.

(Both) User Choice and Control: While nearly every US bank (49, or 98%) provided information about a user's ability to opt out of data collection sharing, only 23 mobile money services (77%) offered information about a user's options to control the use of their data. We note that the United States defines a right of a customer to limit certain types of sharing, and this may explain why US banks discuss this at a greater rate. Nevertheless, this principle is present in both GSMA and FDIC recommendations, so mobile money services should improve their discussion of this issue.

(Both) Security: Both the GSMA and the FDIC recommend that services provide information about the security mechanisms used to protect the collected personal information. Nearly every US bank studied offers this information in a privacy policy (49 policies, or 98%). In addition, 26 of the mobile money services (87%) mention data security in some form. It is important to note that in this study we are simply evaluating whether policies discuss security. Given the substantial difficulty in ensuring data security, any security claims must be subject to healthy skepticism [30, 31].

(Both) Data Minimization and Retention: The final principle we measure is whether policies mention a data minimization policy or a data retention policy. We find that many US banks and mobile money services do not cover this. Only 32 US banks (64%) feature policies that cover this principle, and an even smaller number — 12 mobile money services (40%) — discuss this principle. Given the pervasiveness of data breaches, more banks and mobile money services need to adhere to good data minimization and retention policies and inform their customers of these practices.

Aggregate Analysis: With our analysis of each of the individual principles complete, we are now able to judge the overall state of the two markets that we study.

Figure 4 demonstrates the distribution of overall coverage of privacy polices defined by banks and mobile money services in the form of a histogram. All bank policies have at least 6 principles covered, and 20 banks cover exactly 10 principles. By contrast, mobile money policies tend to be far less complete than US banking policies, and it is clear that in absolute terms too many mobile money services do not have privacy policies that adhere to well-established best practices.

5.3 Readability

Our next goal was to characterize the readability of privacy policies. As discussed in Section 4.4, we use a series of grade-level estimation techniques from the linguistics community to score each policy. Of the 30 mobile money services that had privacy policies, 23 of these were originally written in English. Because different languages have different characteristics in terms of sentence structure and verbosity, to ensure that our results were consistent, we only calculated the readability scores of those 23 policies.

Figure 5 shows our results. US Banks scored an average grade-level readability score of 10.8 ($\sigma^2 = 1.9$), and had a range of between 8.2 (Northern Trust) and 14.6 (Deutsche Bank). Mobile money services had a higher mean reading level of 12.1 with much greater variance ($\sigma^2 = 5.3$), and a range of 9.4 (UseBoom) to 18.4 (Indosat).

To determine if the differences in scores in these two populations are statistically significant, we performed a two-tailed Mann-Whitney-U test. We selected this test over a traditional t-test because it does not assume that the populations are normally distributed. Our null hypothesis was that there is no difference between the readability of the mobile money and traditional banking policies. We selected a significance threshold $\alpha = 0.001$. We note that we chose this extremely conservative threshold to control for the fact that our two datasets differed in variance. The analysis resulted in a z-score of -3.29525 with a corresponding p-value of 0.00096, which is below our conservative threshold of significance. We also calculated an effect size of $r = 0.39$, which represents between a medium (0.3) and large (0.5) effect size [12]. Therefore, the null hypothesis is rejected and there is a statistically significant difference between the readability of privacy policies of mobile money services and U.S. banks. The implication of this analysis is that mobile money policies, on average, appear to be more difficult to comprehend than their traditional banking counterparts.

To further understand the difference between these two sets, we then characterized policy lengths. US Banks had a mean count of 1492 words ($\sigma^2 = 660.7$). State Street's policy had the highest word count (3494), while First Merit was the lowest (557). In general, policies with greater word counts tended to have lower readability scores. Surprisingly, mobile money services had a slightly shorter mean length of 1374 words but with dramatically higher variance ($\sigma^2 = 1373.2$). These results can be better explained by looking at specific data points. Suvidhaa, for instance, had the greatest word count (5518), while EcoCash had the shortest (68). EcoCash is not alone in writing an extremely short policy; TigoPesa's policy is only 268 words long. Figures 8 and 9 show the entirety of these two short policies.

We again sought to determine if our observations were statistically significant. Accordingly, we performed a two-tailed Mann-Whitney-U analysis on the word count results. For that analysis, we again set $\alpha = 0.001$ to control for the increased likelihood of a Type-1 error given the differences in variance of the datasets. Our null hypothesis was that there was no difference in the length of the privacy policies for the mobile money services and US Banks. The analysis yields a z-score of 7.08221 with p-value less than 0.00001. Moreover, with a large calculated effect size of $r = 0.83$, we determined that our results were indeed statistically significant and we could reject the null hypothesis. This implies that length of privacy policies for mobile money services differs significantly from those of traditional banks.

As a final measure of readability, we plotted our measured word counts against the grade-level estimations. Figure 6 shows our results, and includes two important trendlines. While both mobile money and US Banks see the grade-level requirement to understand their policies increase as the word count increases, mobile money services experience this trend in a greatly accelerated fashion. Second, while the privacy community has generally advocated for shorter policies in the past, our readability and coverage analyses demonstrate that shorter policies alone are not necessarily "better." Counterintuitively, mobile money services tend to have short policies that are harder to read.

5.4 Language

In our final analysis, our goal is to determine whether the privacy policies supplied by banks and mobile money services are available in popular or official languages in the countries where they operate. This is a more general question of whether these policies are actually readable by the population for whom they are designed to serve, also noted by [44]. For example, a low-grade level policy written in English is still not readable to a customer who only speaks French. The question of language availability is critical since the principles expressed by these

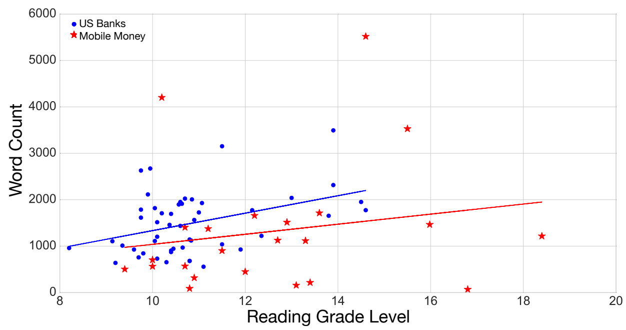


Figure 6: Reading grade level vs. word count of U.S. bank policies and mobile money policies. Note that mobile money policies tend to be shorter and yet more difficult to read.

policies become meaningless if the population of customers is unable to read them.

For US banks, every privacy policy that we examined was available in English. While the United States does not have an official language, English is by far the most popular language of communication, with over 230 million people primarily speaking it at home, while Spanish is the second-most popular language spoken in the country [35]. US banks fall short in addressing the needs of these customers as only 13 of the 50 (or 25%) of financial institutions we reviewed listed privacy policies in Spanish on their website. Furthermore, Spanish-speaking customers would likely have difficulty even accessing this information from these websites, as only two institutions, Fifth Third and East West (4% of the total) allowed users to view the entire website in Spanish.

While US banks could make greater efforts to make their privacy policies accessible to speakers of other languages, the challenges faced in providing policies in readable languages in the US pale in comparison to those presented in the mobile money space. As discussed earlier in this section, only 30 of the 54 mobile money services that we examined had any privacy policies at all. We used the CIA World Factbook [6] to determine the most widely-spoken languages in the countries of operation for these services. Many of the countries where these mobile money services are deployed have more than one official or commonly spoken language, making the question of language accessibility even more important.

Figure 7 shows the extent of these language accessibility challenges for mobile money services. We found that of the 30 services with policies, only 20 were available in either the first or second most widely-spoken language of the service's native country, with 16 policies available in the most widely-spoken language and 6 available in the second-most widely-spoken language. This means that 10 of the 30 mobile money services with any sort of privacy policy, or 33% of this group, *do not have policies written in languages readable by speakers of the most widely-used languages in their countries.* These populations are disempowered from learning about their privacy rights because of this lack of language accessibility.

We also found that only 13 mobile money service websites (or 43% of the services with a privacy policy) were available in either the first or second-most widely spoken language within that country.

These issues represent a serious impediment to inclusion and privacy. The lack of accessible material results in the inability for large segments of the population to be able to make informed choices regarding their privacy. It is imperative that customers have the opportunity to understand their rights and options for controlling their personal data, and there can be no meaningful ability to do so unless customers are provided materials comprehensible to them in the languages they use.

5.5 Mobile Payment Apps

Our previous results showed a clear difference between privacy policies in US Banks and mobile money systems. We were also interested to know if popular mobile payment apps from developed countries performed well according to our criteria of coverage, readability, and language availability. To that end, we look at the two most popular mobile payment apps in the US: Paypal and Venmo. We found that policies covered all 11 principles with only a few exceptions: Paypal had no coverage of children's policy, while Venmo did not cover data minimization or retention. Neither policy covered employee accountability or enforcement. Venmo's average reading grade level was 13.2, while Paypal scored higher at 14.9. In addition, Paypal's word count of 3,239 was over 1000 more words than Venmo's policy (2,065 words). Although both mobile payment apps are used widely across the US, we were unable to find a privacy policy in any other language that English. We note that our sample size of 2 apps means meaningful statistical analysis is simply not possible. However, these results are similar to our findings for US banks.

6. IMPROVING MOBILE MONEY PRIVACY DISCLOSURES

Our results show that the mobile money industry, as a whole, does not provide sufficient disclosure of privacy practices. The question then is how can we improve the state of privacy disclosure in mobile money? In this subsection, we discuss the role that regulation by national governments as well as industry-driven "self-regulation" may play in improving this state of affairs. We also discuss what future improvements to mobile money privacy policy recommendations should entail. We note that while we strongly believe improvement is imperative for mobile money privacy disclosures and privacy practices, we do not take a strong stance on which path is best. Finally, we conclude by acknowledging that norms for privacy vary from culture to culture.

6.1 Regulating Mobile Money Policies

One possible path to improving mobile money privacy policies is legal regulations in the spirit of the Gramm-Leach-Bliley Act (GLBA) in the United States. Prior work by Sheng and Cranor [37] showed that the GLBA significantly improved the coverage of privacy disclo-

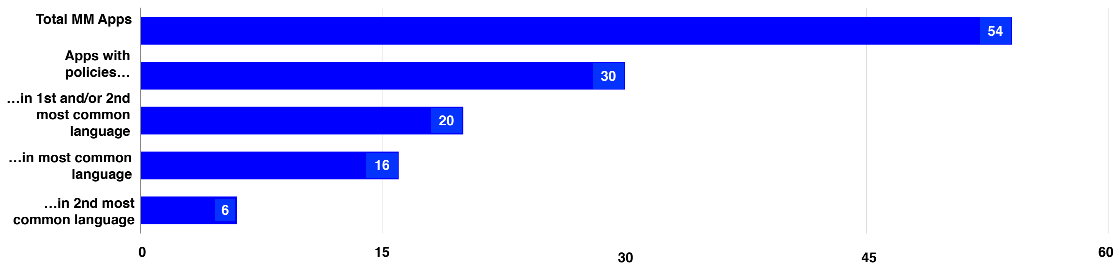


Figure 7: Language analysis of Mobile Money service policies. Note that 18% of all services (54) and 33% of services that actually have privacy policies (30) are neither available in the first or second official languages of the country in which they are deployed.

sures in the US. The data that we present on US policies confirms this earlier finding: US policies are largely complete according to the FDIC standards. We do note, however, that regulation is not a panacea. For example, some US policies lacked coverage of important information. More importantly, regulations about *disclosure* do not necessarily minimize data collection or mean that users have meaningful options about that collection [37].

We note that we are among the first to look at privacy policies from an international perspective. This comes with a number of difficulties. For example, each country will have its own laws, regulations, and local operating practices. In mobile money this is complicated by the fact that these services can be regulated by many (sometimes overlapping) agencies. These agencies include financial regulators, telecommunications regulators, or others [41]. In future work we hope to partner with experts in law in various countries served by mobile money to determine what, if any existing laws or regulations could apply to mobile money, and how effective those mechanisms are. As a case study example of the role of regulation, we examined India's privacy regulations for all businesses (not specific to mobile finance) as well as the performance of mobile money systems based in India. These regulations are specified in a regulation document published by the Ministry of Communications and Information Technology titled the "Information Technology (Reasonable security practices and procedures and sensitive personal data or information) Rules, 2011." [27] This document mandates that organizations subject to the IT Amendment Act of 2008 must provide privacy notices to users. The regulation also mandates that those notices contain information mapping to four of our analysis principles: "Purpose and Use," "Examples," "Security," and "Sharing Process." We note that the governing regulations for mobile banking by the Reserve Bank of India make no mention of privacy practices [34].

We found that all India-based systems we studied had privacy policies. More importantly we found that their coverage was among the best of mobile money systems that had a policy. All discussed the data collection purpose, what data is collected, definitions of terms, third party sharing, users' options for data practices, concrete examples of types of data, data security, and data retention and minimization policies. Three out of four apps discussed specific third-party data sharing practices, and one service (Oxigen Wallet) discussed employee accountability. No Indian services specifically addressed children's data privacy. In total, Indian services had a higher percentage coverage of every privacy policy principle than mobile money systems from all other countries (with the exception of children's data privacy); they also met the criteria required by the IT law. This small case study does not causally prove that regulation leads to better policies, but it does motivate further exploration of this idea.

Change in mobile money as a result of national regulations will likely take time, and this is complicated by several factors. For example, it is not always clear which entity in a country is responsible for setting and enforcing policy related to mobile money. Because this industry does not fall under the same regulatory environment as traditional banking, authority is often scattered across multiple parts of a government (e.g., the telecommunications bureau, the central bank, etc.) [41]. Government-enforced regulations may be made easier through the upcoming creation of transnational unions such as the Economic Community of West African States (ECOWAS), which plans to share a single currency and set of policies among its 15 member nations by 2020 [14].

Another route to improve privacy disclosures in mobile money would be through "self-regulation" through an industry consortium like the GSMA, whose recommendations we use in this paper [43]. Advocates of this approach argue that those within the industry are best equipped to determine what users need and balance it with the needs of that industry. Such "self-regulation" may not provide the same enforceable guarantees as government-enforced regulations. In particular, industry groups must strike a balance between representing their member companies' interests and in requiring said companies to change their practices. "Self-regulation" does have some advantages though. For example, industry associations can deploy recommendations faster than is typically possible for governments, and it can standardize these recommendations internationally. We note that industry-based guidance is not necessarily mutually exclusive of national regulation.

6.2 Recommendations for Privacy Policy Standards

Orthogonal to the question of how to induce change in mobile money services is specifying "in what ways should these policies change."

Expanding Policy Coverage: We believe that the GSMA standards are a minimal starting point, but recognize they are deficient in a number of ways. First, they are woefully underspecified. The principal document defining them could be characterized as an "infographic." Second, the coverage should be expanded to cover additional areas of concern. These include the areas of coverage in the FDIC standards: describing the collection process, providing definitions of important terms, providing clear, concrete examples of how data will be used, and how data will be shared with third parties. As we discussed earlier, we note that the FDIC standards are not necessarily ideal, but they provide a strong starting point for determining a complete privacy policy. Other standards and guidelines from governments or consumer protection organizations — including established privacy policy templates and generating tools [2–4] — may also be instructive for future mobile money privacy policy standards.

Expanding User Comprehension: Finally, we note that for privacy policies to have value, users must be able to understand how their privacy is affected by using these services so they can make informed decisions. As our analysis demonstrates, privacy policies for these services often lack content and are written in ways that impair readability. In many cases, significant populations cannot make decisions as policies are not written in languages they understand. It is therefore vital that these policies are not only complete, but written in ways that allow users to understand how their data is used.

This issue is further amplified by literacy rates that can vary widely between countries. For example, Qatar's literacy rate is over 97% while the literacy rate in Mali is less than 47% [6]. However, none of the privacy policies that we examined considered how to effectively communicate policy details to illiterate customers. In countries where literacy rates are low, it is important to consider new ways of making mobile money customers aware of their privacy rights.

6.3 Cultural Norms for Privacy

Deciding how privacy should be protected across a set of services that span a wide array of cultures and continents was not a simple task. In many parts of the world, especially Europe, privacy is carefully guarded and assumed to be a human right. Chinese culture, however, instead often values privacy less when compared to community, order and governance [24]. Similarly, in settings where sharing or communal ownership (e.g., of cell phones) are common, there are different standards for individual privacy [29].

Accordingly, our selection of the GSMA policies was made only after careful consideration. In addition, the GSMA claims to be the embodiment international understanding on privacy. Instead of attempting to pick a universal set of values for privacy across mobile money services, we felt that the best available consensus on the matter likely comes from the industry itself. That is not to say that the protections provided by the GSMA, FDIC or any currently available policy are perfect. Rather, they form the only available lenses through which we can observe the current state of global privacy expectations in the digital financial services space.

We believe that significant work remains to be done in this space. As efforts towards interoperable services increase [33], questions about which country's privacy rules dominate in cross-border transactions remain unanswered. Moreover, methods of communicating such policies to users whose cultural frame of reference and literacy may vary widely will also prove challenging.

7. CONCLUSION

Mobile money services provide new abilities for customers to use their mobile phones to make payments, significantly broadening financial inclusion and helping to raise people out of poverty. However, the privacy guarantees of these services has remained unexplored. We conducted a comprehensive analysis on privacy policies of all 54 mobile money services that provided smartphone apps, and compared these policies to the top 50 US banks by assets. We found that although all US banks had privacy policies, over 44% of mobile money services had no privacy policy whatsoever. For those services that did have policies, most were missing key factors, including privacy principles laid out by industry groups that these services agreed to uphold. Moreover, compared to bank policies, mobile money policies were hampered by being difficult to read, even though they were on average significantly shorter. Several mobile money services did not even offer policies written in the languages used by a majorities of their target population. Our study represents a call to action for operators, governments, and NGOs, to assure that agreed-upon principles and policies are enforced, expanded upon, and made accessible to the customers of these services in order to better protect their privacy.

Acknowledgments

We thank Jami Solli for her insights. We also thank Blase Ur, Matthew Smith, and the anonymous reviewers of our work for their valuable comments. This work was supported in part by the National Science Foundation under grant numbers CNS-1526718, and CNS-1540217. 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.

8. REFERENCES

[1] Ecocash - android apps on google play. https://play.google.com/store/apps/details?id=com.econet.ecocash.

[2] FreePrivacyPolicy.com.

[3] PrivacyPolicies.com.

[4] Final Model Privacy Form Under the Gramm-Leach-Bliley Act - 16 CFR Part 313. http://bit.ly/1cQG1ya, December 2009.

[5] Federal Reserve Statistical Release: Large Commercial Banks. https://www.federalreserve.gov/releases/lbr/current/, September 2016.

[6] The CIA World Factbook. https://www.cia.gov/library/publications/the-world-factbook/geos/, January 2017.

[7] Asli Demirguc-Kunt, Leora Klapper, Dorothe Singer, and Peter Van Oudheusden. The Global Findex Database 2014 Measuring Financial Inclusion around the World. Technical Report Policy Research Working Paper 7255, World Bank Group, April 2015.

[8] S. Badarudeen and S. Sabharwal. Assessing Readability of Patient Education Materials: Current Role in Orthopaedics. *Clinical Orthopaedics and Related Research*, 468(10):2572–2580, October 2010.

[9] E. Beaunoyer, M. Arsenault, A. M. Lomanowska, and M. J. Guitton. Understanding online health information: Evaluation, tools, and strategies. *Patient Education and Counseling*, 100(2):183 – 189, 2017.

[10] K. Butler, L. Perlman, P. Makin, H. Gerwitz, P. Traynor, Y. Grin, E. Bondarenko, and R. Miller. ITU-T Focus Group Digital Financial Services: Security Aspects of Digital Financial Services (DFS). Technical report, International Telecommunications Union Standardization Sector (ITU-T), December 2016.

[11] S. Castle, F. Pervaiz, G. Weld, F. Roesner, and R. Anderson. Let's Talk Money: Evaluating the Security Challenges of Mobile Money in the Developing World. In *7th ACM Symposium on Computing for Development (DEV)*, November 2016.

[12] J. Cohen. A Power Primer. *Psychological Bulletin*, (1):155–159, July 1992.

[13] L. F. Cranor, P. G. Leon, and B. Ur. A Large-Scale Evaluation of U.S. Financial Institutions' Standardized Privacy Notices. *ACM Trans. Web*, 10(3):17:1–17:33, August 2016.

[14] Economic Community of West African States (ECOWAS). ECOWAS VISION 2020: Towards A Democratic And Prosperous Community. http://www.ecowas.int/wp-content/uploads/2015/01/ECOWAS-VISION-2020.pdf, 2010.

[15] J. A. Eloy, S. Li, K. Kasabwala, N. Agarwal, D. R. Hansberry, S. Baredes, and M. Setzen. Readability assessment of patient education materials on major otolaryngology association websites. *Otolaryngology – Head and Neck Surgery*, 147(5):848–854, 2012.

[16] Federal Deposit Insurance Corporation. Privacy Rule Handbook. https://www.fdic.gov/regulations/examinations/financialprivacy/handbook/, 2001.

[17] J. Gluck, F. Schaub, A. Friedman, H. Habib, N. Sadeh, L. F. Cranor, and Y. Agarwal. How Short Is Too Short? Implications of Length and Framing on the Effectiveness of Privacy Notices. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 321–340, Denver, CO, June 2016. USENIX Association.

[18] MMU Deployment Tracker. http://www.gsma.com/mobilefordevelopment/programmes/mobile-money-for-the-unbanked/insights/tracker, November 2016.

[19] M. Hochhauser. Lost in the Fine Print: Readability of Financial Privacy Notices. https://www.privacyrights.org/blog/lost-fine-print-readability-financial-privacy-notices-hochhauser, July 2001.

[20] C. Jensen and C. 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*, April 2004.

[21] M. Johnson, J. Karat, C.-M. Karat, and K. Grueneberg. Optimizing a Policy Authoring Framework for Security and Privacy Policies. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, July 2010.

[22] P. G. Kelley, J. Bresee, L. F. Cranor, and R. W. Reeder. A "Nutrition Label" for Privacy. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*, July 2009.

[23] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[24] T. Li and Z. Zhou. Do You Care About Chinese Privacy Law? Well, You Should. https://iapp.org/news/a/do-you-care-about-chinese-privacy-law-well-you-should/, January 2015.

[25] A. M. McDonald and L. F. Cranor. The Cost of Reading Privacy Policies. *I/S: A Journal of Law and Policy for the Information Society*, 4(3):540–565, 2008.

[26] C. Mims. 31% of Kenya's GDP is Spent Through Mobile Phones. http://qz.com/57504/31-of-kenyas-gdp-is-spent-through-mobile-phones/, February 2013.

[27] Ministry of Communications and Information Technology of India. Information Technology (Reasonable security practices and procedures and sensitive personal data or information) Rules, 2011. http://www.wipo.int/edocs/lexdocs/laws/en/in/in098en.pdf, 2011.

[28] M. Paik. Stragglers of the Herd Get Eaten: Security Concerns for GSM Mobile Banking

Applications. In *Proceedings of the Workshop on Mobile Computer Systems and Applications (HotMobile)*. ACM, February 2010.

[29] J. Poushter, J. Bell, D. Cuddington, K. Devlin, M. Keegan, B. Parker, K. S. B. Stokes, R. Wike, and H. Zainulbhai. Cell Phones in Africa: Communication Lifeline Texting Most Common Activity, but Mobile Money Popular in Several Countries. Technical report, Pew Research Center, April 2015.

[30] B. Reaves, J. Bowers, N. Scaife, A. Bates, A. Bharatiya, P. Traynor, and K. Butler. Mo(bile) Money, Mo(bile) Problems: Analysis of Branchless Banking Applications in the Developing World. *ACM Transactions on Privacy and Security (TOPS)*, 2017.

[31] B. Reaves, N. Scaife, A. Bates, P. Traynor, and K. Butler. Mo(bile) Money, Mo(bile) Problems: Analysis of Branchless Banking Applications in the Developing World. In *Proceedings of the USENIX Security Symposium (SECURITY)*, August 2015.

[32] R. W. Reeder, C.-M. Karat, J. Karat, and C. Brodie. Usability Challenges in Security and Privacy Policy-authoring Interfaces. In *Proceedings of the International Conference on Human-computer Interaction*, September 2007.

[33] D. G. Reiss and R. T. L. Mourao. The Regulator's Perspective on the Right Timing for Inducing Interoperability - Findings of a survey among Focus Group Members. Technical report, International Telecommunications Union Standardization Sector (ITU-T), February 2017.

[34] Reserve Bank of India. Master Circular - KYC norms, AML standards, CFT, Obligation of banks under PMLA, 2002. `http://rbidocs.rbi.org.in/rdocs/notification/PDFs/94CF010713FL.pdf`, 2013.

[35] C. Ryan. Language Use in the United States: 2011 – American Community Survey Reports. United States Census Bureau. `https://www.census.gov/prod/2013pubs/acs-22.pdf`, August 2013.

[36] F. Schaub, R. Balebako, A. L. Durity, and L. F. Cranor. A Design Space for Effective Privacy Notices. In *Symposium On Usable Privacy and Security (SOUPS)*, July 2015.

[37] X. Sheng and L. F. Cranor. Evaluation of the Effect of US Financial Privacy Legislation Through the Analysis of Privacy Policies. *ISJLP*, 2:943, 2005.

[38] W. Shepard. After Day 50: The Results From India's Demonetization Campaign Are In. `https://www.forbes.com/sites/wadeshepard/2017/01/03/after-day-50-the-results-from-indias-demonetization-campaign-are-in/`, January 2017.

[39] T. Suri and W. Jack. The Long-Run Poverty and Gender Impacts of Mobile mMoney. *Science*, 354(6317):1288–1292, December 2016.

[40] P. F. Svider, N. Agarwal, O. J. Choudhry, A. F. Hajart, S. Baredes, J. K. Liu, and J. A. Eloy. Readability assessment of online patient education materials from academic otolaryngology – head and neck surgery departments. *American Journal of Otolaryngology*, 34(1):31 – 35, 2013.

[41] N. A. Tagoe. Who Regulates the Mobile Money Operations by Telco's? The Need for an Effective and Robust Legislative and Regulatory Framework in Ghana. *Journal of Business and Financial Affairs*, 5(3), August 2016.

[42] The GSM Association (GSMA). Code of Conduct for Mobile Money Providers, Version 2. `http://www.gsma.com/mobilefordevelopment/wp-content/uploads/2015/10/Code-of-Conduct-for-Mobile-Money-Providers-V2.pdf`, 2015.

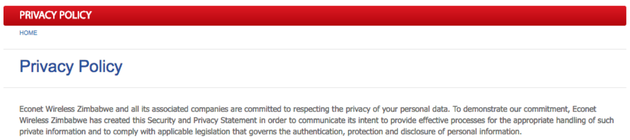
[43] The GSM Association (GSMA). Mobile Privacy Principles: Promoting Consumer Privacy in the Mobile Ecosystem. `http://www.gsma.com/publicpolicy/wp-content/uploads/2012/03/GSMA2016_Guidelines_Mobile_Privacy_Principles.pdf`, 2016.

[44] B. Ur, M. Sleeper, and L. F. Cranor. {Privacy, Privacidad, Приватност} policies in social media: Providing translated privacy notice. *I/S: A Journal of Law and Policy for the Information Society*, 9(2), 2013.

[45] US Congress. Gramm-Leach-Bliley Act, Financial Privacy Rule. `https://www.gpo.gov/fdsys/pkg/PLAW-106publ102/content-detail.html`, November 1999.

[46] S. Zimmeck and S. M. Bellovin. Privee: An Architecture for Automatically Analyzing Web Privacy Policies. In *23rd USENIX Security Symposium (USENIX Security 14)*, pages 1–16, San Diego, CA, August 2014. USENIX Association.

APPENDIX



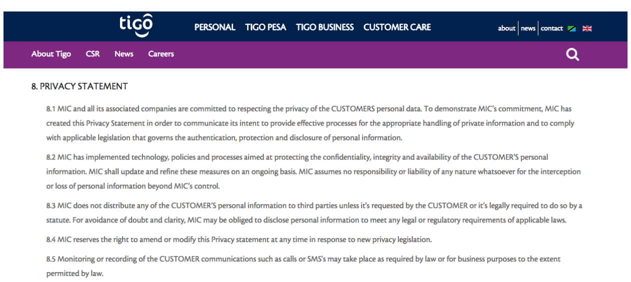
PRIVACY POLICY

HOME

Privacy Policy

Econet Wireless Zimbabwe and all its associated companies are committed to respecting the privacy of your personal data. To demonstrate our commitment, Econet Wireless Zimbabwe has created this Security and Privacy Statement in order to communicate its intent to provide effective processes for the appropriate handling of such private information and to comply with applicable legislation that governs the authentication, protection and disclosure of personal information.

Figure 8: EcoCash, which has endorsed the GSMA's Code of Conduct, has a very short (68 word) privacy policy. In its current state, it only meets one of the GSMA's recommendations.



tigo PERSONAL TIGO PESA TIGO BUSINESS CUSTOMER CARE about | news | contact

About Tigo CSR News Careers

8. PRIVACY STATEMENT

8.1 MIC and all its associated companies are committed to respecting the privacy of the CUSTOMERS personal data. To demonstrate MIC's commitment, MIC has created this Privacy Statement in order to communicate its intent to provide effective processes for the appropriate handling of private information and to comply with applicable legislation that governs the authentication, protection and disclosure of personal information.

8.2 MIC has implemented technology, policies and processes aimed at protecting the confidentiality, integrity and availability of the CUSTOMER'S personal information. MIC shall update and refine these measures on an ongoing basis. MIC assumes no responsibility or liability of any nature whatsoever for the interception or loss of personal information beyond MIC's control.

8.3 MIC does not distribute any of the CUSTOMER'S personal information to third parties unless it's requested by the CUSTOMER or it's legally required to do so by a statute. For avoidance of doubt and clarity, MIC may be obliged to disclose personal information to meet any legal or regulatory requirements of applicable laws.

8.4 MIC reserves the right to amend or modify this Privacy statement at any time in response to new privacy legislation.

8.5 Monitoring or recording of the CUSTOMER communications such as calls or SMS's may take place as required by law or for business purposes to the extent permitted by law.

Figure 9: TigoPesa (Tanzania)'s short privacy policy (268 words) is a subsection within TigoPesa's Terms and Conditions. In its current state, it only meets one GSMA principle and one FDIC principle.

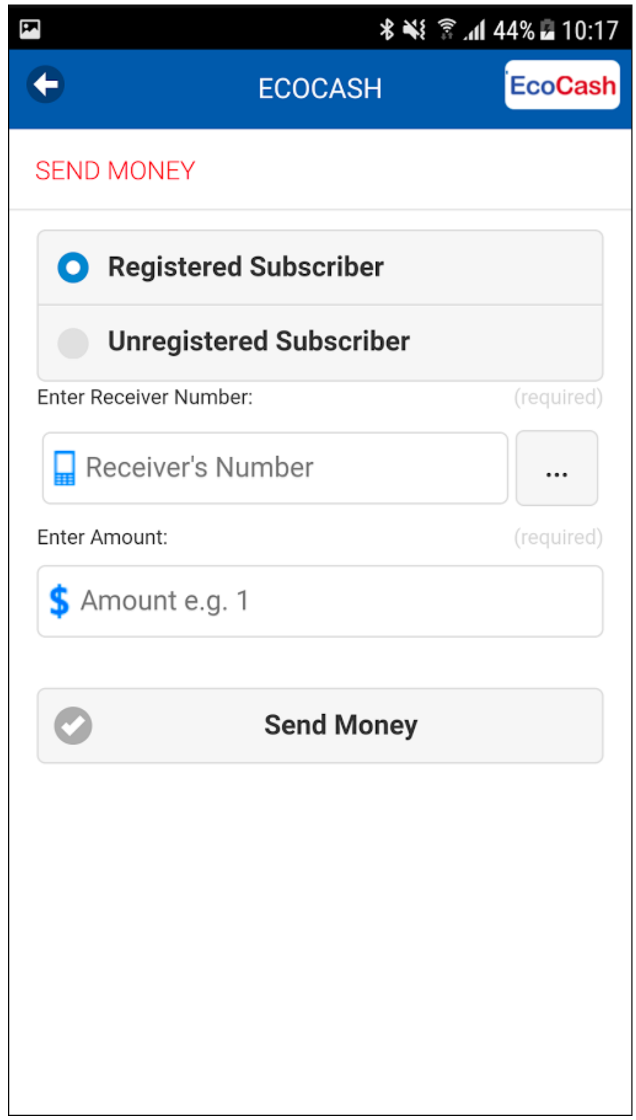


Figure 10: EcoCash mobile money application screenshot (obtained from the Google Play Store [1]).

Table 1: US Policies: Principles Included

	GSMA			FDIC				FDIC and GSMA			
Bank Name	Purpose of Data Collection	Children and Adolescents	Accountability and Enforcement	Collection Process	Definitions	Examples	Third Parties	User Choice and Control	Security	Sharing Process	Data Minimization/Retention
Ally	✓	✓		✓	✓	✓	✓		✓	✓	
Associated	✓			✓	✓	✓		✓	✓	✓	✓
Bank of America	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
Bank of NY	✓			✓		✓	✓	✓		✓	
Bank of West	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
Bank United	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Barclays	✓			✓	✓	✓	✓	✓	✓	✓	✓
BB&T	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
BBVA	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
BMO	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
BOK Financial	✓			✓	✓	✓		✓	✓	✓	✓
Capital One	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Chase	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
CIT	✓	✓		✓	✓	✓	✓	✓	✓	✓	
Citi Bank	✓			✓	✓	✓	✓	✓	✓	✓	
Citizens	✓	✓		✓	✓	✓	✓	✓	✓	✓	
Comerica	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
Deutsche Bank	✓			✓	✓	✓	✓	✓	✓	✓	
Discover	✓	✓		✓	✓	✓	✓	✓	✓	✓	
EastWest	✓		✓	✓	✓	✓		✓	✓	✓	
Fifth Third	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
First Citizens	✓			✓	✓	✓	✓	✓	✓	✓	✓
First Merit	✓			✓	✓	✓		✓	✓	✓	
First Republic	✓			✓	✓	✓	✓	✓	✓	✓	
First Tennessee	✓			✓	✓	✓		✓	✓	✓	✓
Frost Bank	✓	✓		✓	✓	✓	✓	✓	✓	✓	
Goldman Sachs	✓			✓	✓	✓	✓	✓	✓	✓	
HSBC	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
Huntington	✓	✓		✓	✓	✓	✓	✓	✓	✓	
JP Morgan	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
Key Corp	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
M&T	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
Zions	✓			✓	✓	✓		✓	✓	✓	✓
Wells Fargo	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
Webster	✓			✓	✓	✓		✓	✓	✓	✓
US Bank	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
Union Bank	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
TD Bank	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
Synovus	✓			✓	✓	✓	✓	✓	✓	✓	✓
Sun Trust	✓			✓	✓	✓	✓	✓	✓	✓	✓
State Street Bank	✓			✓	✓	✓	✓	✓	✓	✓	
Silicon Valley	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
Santander	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
RBC	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Regions	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
PNC	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
Peoples	✓			✓	✓	✓	✓	✓	✓	✓	✓
Northern Trust	✓			✓	✓	✓		✓	✓	✓	✓
Morgan Stanley	✓			✓	✓	✓	✓	✓	✓	✓	
Signature	✓			✓	✓	✓		✓	✓	✓	✓

Table 2: Mobile Money Policies: Principles Included

		GSMA			FDIC				FDIC and GSMA			
Service Name	Country	Purpose of Data Collection	Children and Adolescents	Accountability and Enforcement	Collection Process	Definitions	Examples	Third Parties	User Choice and Control	Security	Sharing Process	Data Minimization/Retention
Airtel Money	India	✓			✓	✓	✓	✓	✓	✓		✓
Bits	Uruguay				✓					✓		
EcoCash	Zimbabwe									✓		
eSewa	Nepal	✓			✓	✓		✓	✓	✓	✓	✓
EZcash	Sri Lanka	✓			✓	✓	✓	✓	✓	✓		
FNB	South Africa	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
GCash	Phillipines	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
IdeaMyCash	India	✓			✓	✓	✓	✓	✓	✓	✓	✓
Indosat	Indoneisa	✓			✓		✓	✓		✓	✓	
Mcash	Singapore	✓			✓		✓	✓	✓	✓	✓	✓
mCoin	Indonesia	✓			✓	✓	✓	✓	✓	✓	✓	
Mdinar	Tunisia								✓	✓		
MobiCash	Mali	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
Mobile Money NG	Nigeria	✓	✓		✓			✓	✓	✓	✓	
mPay	Thailand	✓			✓	✓	✓		✓	✓	✓	✓
Ooredoo	Qatar	✓	✓		✓			✓	✓	✓	✓	
Orange Money	Côte d'Ivoire	✓			✓				✓			
Oxigen	India	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓
Paga	Nigeria	✓		✓	✓			✓	✓	✓	✓	✓
Simba	Lebanon									✓		
Standard Bank	South Africa	✓			✓	✓		✓	✓	✓	✓	
Suvidhaa Money	India	✓			✓	✓	✓	✓	✓	✓	✓	✓
Teasy Mobile	Nigeria				✓			✓	✓	✓	✓	✓
Tigo Pesa	Tanzania							✓			✓	
Tigo SV	El Savador							✓	✓			
TPago	Dominican Republic					✓		✓		✓		
True Money	Thailand				✓			✓	✓		✓	
UseBoom	Mexico	✓			✓	✓	✓		✓	✓	✓	
Zenith	Nigeria				✓		✓	✓	✓	✓	✓	
Zuum	Brazil				✓			✓		✓	✓	
No Policies Available:												
BKash	Bangladesh											
Ecash	Indonesia											
eZuza	South Africa											
First Monie	Nigeria											
Fortis Mobile Money	Nigeria											
Growth Enhancement Support Scheme	Nigeria											
JCUES Mobile Money	Jamaica											
Mi Billetera Movil	Argentina											
Mobile Money Guyana	Guyana											
MoneyOnMobile	India											
mService	Vietnam											
mVola	Madagascar											
my.wallet	Nigeria											
Myanmar Mobile Money	Myanmar											
Oi Carteira	Brazil											
Pido	Nigeria											
Pocket Moni	Nigeria											
Qash Mobile Banking	Côte d'Ivoire											
Ready Cash	Nigeria											
Splash Cash	Sierra Leone											
Tigo Honduras	Hondurus											
Tigo Money Bolivia	Bolivia											
VCash	Nigeria											
Wizzit	South Africa											

Table 3: Keywords and Phrases

Principle	Key Words and Phrases
Purpose of Data Collection	Reasoning, Enhance User Experience, User Experience
Children and Adolescents	Children, Children's Privacy
Accountability and Enforcement	Employee, Accountable, Accountability
Collection Process	Collect
Definitions	Means, Is, Are
Examples	Types of Personal Information, Types Of, For Example, Includes
Third Parties	Third Party, Third Parties
User Choice and Control	Disable, Edit, User Can, Change
Security	Security
Sharing Process	Share, Sharing Process
Data Minimization and Retention	Minimization, Termination, Continue to share, Retention, Retain

A Qualitative Investigation of Bank Employee Experiences of Information Security and Phishing

Dan Conway, Ronnie Taib, Mitch Harris, Shlomo Berkovsky, Kun Yu, Fang Chen

Data61, CSIRO, Australia
firstname.lastname@data61.csiro.au

ABSTRACT

Staff behaviour is increasingly understood to be an important determinant of an organisations' vulnerability to information security breaches. In parallel to the HCI and CSCW literature, models drawn from cognitive and health psychology have suggested a number of mental variables that predict staff response to security threats. This study began with these models, but engaged in a broader, discovery-orientated, qualitative investigation of how these variables were experienced, interacted subjectively, and what further variables might be of relevance. We conducted in-depth, semi-structured interviews consisting of open and closed questions with staff from a financial services institution under conditions of strict anonymity. Results include a number of findings such as a possible association between highly visible security procedures and low perceptions of vulnerability leading to poor security practices. We also found self-efficacy was a strong determinant of staff sharing stories of negative experiences and variances in the number of non-relevant emails that they process. These findings lead to a richer, deeper understanding of staff experiences in relation to information security and phishing.

1. INTRODUCTION

The roles that staff play in information security (IS) breaches have, of late, become increasingly recognised as important determinants of an organisation's IS defence posture. While the exact classification of breach types remains controversial, reports such as IBM's 2014 Cyber Security Intelligence Index claim that 'human error' was implicated in over 95% of significant data breaches of their systems [1]. As such, it is becoming apparent that purely technical solutions to information security will not be sufficient to address the growing threat to our networks and data posed by cyber criminals and hostile entities.

There are a number of much discussed user failures to comply with IS policies that have shown to be largely explicable using investigations based around user-education and the usability-security trade-off. Examples include; the difficulties in complying with password policies [2, 3], giving away too much personal information when not required [4], and ignoring warning messages when engaging in unsafe behaviour [5]. As these examples suggest, this body of work is typically based on 'user studies' where the dependent variables are either behavioural, or subjective observations of the behaviours in question. This body of literature also typically focusses on raising user awareness of cyber threats, with the assumption that knowledge will allow people to recognise and deal with attacks. However, a further class of problems requires a different investigative lens. Cyber attackers are now recognised as understanding and leveraging the inherent cognitive biases and weaknesses of the human information processing system [6, 7], enabling them to bypass effortful, deep information processing by the user [8]. This is particularly evident in phishing attacks, which consist of generic, non-targeted emails, distributed widely, that attempt to entice the user to click on a link or open an attachment leading to a malware infection or security credentials being revealed to the attacker. These types of exploits are crafted with increasing sophistication aimed at bypassing conscious processing of the victim and eliciting more automatic behaviours characterised by shallow information processing and as such these methods require new approaches to mitigate [9]. In the face of these kinds of attacks, analyses based on more behavioural methods are likely to fall short, explication requiring a deeper engagement with the cognitive processes that staff experience when facing threats. In this paper we discuss cognitive models that include constructs such as threat Self-Efficacy (SE) and perceived Vulnerability (V). These variables, in particular, have been shown to predict users deploying protective behaviours to a greater extent than knowledge alone [10, 11]. Knowledge is now seen as necessary, but not sufficient to arm users against attackers.

This paper aims to extend the understanding of the human end-user within the IS landscape, specifically seeking to understand the underlying, presumably causative, cognitive variables that drive these behaviours. This work draws on the literature of cognitive psychology and aims to extend the approaches adopted by the HCI and CSCW community. Our study involved the staff of a major financial services institution in Australia and New Zealand. The study was aimed principally at understanding factors implicated in victimisation via phishing attacks, but had as a secondary objective to understand the challenges that staff faced in relation to IS more generally. We were interested in the following research questions:

- What cognitive variables may be implicated in staff's behaviour in relation to phishing emails?
- How do staff experience information security within the organisation and how does this differ from their perceptions at home?
- What environmental and organisational factors affect staff behaviour in relation to phishing attacks and information security more generally?

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12 -- 14, 2017, Santa Clara, California.

After carrying out our study and analysing the results, we established a number of emergent themes from the data. These themes are detailed individually in the results section of this work and then the implications are considered and additional context is provided in the discussion section. Many of the themes, such as staff's low feelings of vulnerability, variable proportions of non-relevant emails and willingness to share victimisation experiences only if they have high self-efficacy, immediately suggest further working hypotheses, the primary of which are discussed in the future work section.

2. RELATED WORK

Differing approaches to research in cyber security have resulted in subtly, but fundamentally different bodies of literature around the subject. Each has its own characteristics such as assumptions, methods, and investigative lenses. One body of work, emerging largely from the HCI and CSCW domain, has provided us with a rich picture of the behavioural characteristics of users in response to IS challenges. Acquisiti et al. [4] provided an excellent overview of the way users make poor decisions about privacy. Dhamaja et al. [12] demonstrated the inability of people to detect well-crafted phishing emails, even in ideal conditions, and noted the poor response to security indicators such as status bar warnings. And finally users have also been shown to frequently disclose more information on-line than they need to [13], and are often willing to sacrifice privacy for remarkably small rewards [14]. Overall the picture built up by this research is concerning since it indicates that people are extremely vulnerable to cyber attacks.

Much, but not all, of this work is based on an underlying assumption that educating the user will fix the problem. The core issue is often understood to be 'how do we help users learn more about security so that they can make better decisions'. Influential papers such as that by Kumaraguru et al. [15] are focussed almost entirely on the education issue and the dependent variables of interest in the study are all based around the acquisition, retention, and transfer of knowledge.

Arising from this standpoint, an entire commercial ecosystem has emerged offering to address the 'human problem' of cyber security purely through training and education campaigns. However these approaches are rapidly approaching the point of diminishing returns, where security professionals are frustrated by the persistence of poor user IS behaviour leading some authors to suggest that human-based solutions are not feasible and that technical solutions are the only way to effectively safeguard systems from attackers [16].

However, another body of cyber security research has concerned itself more with understanding the underlying cognitive variables, or mental constructs, underpinning the behaviours of interest. This vein of research has its roots in both health psychology and cognitive psychology and promises to extend the efficacy of mitigation methods beyond that offered by simple education. For example, Samaya et al. [10] recently showed, in an excellently designed study of 3,500 participants across seven countries, that user self-confidence in being able to respond to security threats was a more than four times larger predictor of their measure of good cyber security behaviour, than was knowledge of cyber security threats. Findings such as this, that identify the cognitive constructs that drive behavioural models, promise to be able to extend the effectiveness of mitigation strategies beyond the limitations of current 'education and training' approaches.

Models originally based in health psychology [17] are remarkably suited for deployment in the IS domain since the environments are in many ways analogous. Both IS and health involve individual behaviour, in situations of uncertainty and in response to threats which are often poorly understood, where costs can often be temporally far removed and not deemed likely, and where compliance with desired protocols (often referred to as response costs) is either arduous, or not immediately desirable. According to these models, the challenge of eating well and exercising regularly is almost perfectly analogous to deploying strong and different passwords on every system you use.

Many of these current psychological models of behaviour in response to threats are derived to some degree from the hugely influential Theory of Reasoned Action (TRA) [18]. This model proposes that beliefs about a behaviour and evaluations of the outcome of a behaviour result in attitudes towards the behaviour, and that social influences and motivation result in subjective norms. These two constructs; attitude towards the behaviour and subjective norms, then interact to result in behavioural intention, which in turn predicts the behaviour itself. Ajzen's later reformulation of the model into the Theory of Planned Behaviour (TPB) [19] involved the addition of variables that accounted for a person's own beliefs about their ability to carry out the behaviours in question. Referred to as perceived behavioural control and later disambiguated further to variables such as locus of control [20], self-efficacy [17, 21] and response-efficacy [22], these variables have a long history of being shown to be significant predictors of behavioural intentions. These variables are also deployed in contemporary models of protective behaviour such as the Protection Motivation Theory (PMT) [23]. In short, the constructs encompassed in perceived behavioural control suggest that people are unlikely to attempt to engage in a behaviour if they think that they will not be able to carry out the behaviour in question. Constructs such as these are central to our investigations and are discussed in more detail in the discussion section of this paper.

Another prominent model emerging from health psychology literature is Rogers' Protection Motivation Theory (PMT) [11] which was derived from ideas about people's response to fear, and suggested that encountering a communication that induced fear would induce a threat appraisal process which, mediated by variables such as response efficacy, self-efficacy and response costs, would result in protection motivation leading to either an adaptive or maladaptive response to the threat.

More recently still, in the cognitive domain, dual route models of information processing such as the Heuristic Systematic Model (HSM) have begun to be applied specifically to the problem of phishing victimisation with notable successes in predicting user behaviour [9, 24]. Dual Route models suggest that users often engage in little elaborative, deep ('system 2') information processing when scanning emails, and rely instead of more shallow (system 1) information processing based on simple heuristics such as calls to authority, urgency cues and social proof [25] to make fast decisions about whether to respond or not [8]. These models suggest that regardless of a users' knowledge of threats, when scanning an inbox for messages to respond to, users often engage in very shallow cognitive processing of email cues such as sender and subject line, meaning that they are not deploying the knowledge that they have about these cues. This results in important signals such as malformed email addresses (type-jacking) escaping attention. According to these theories, in this low level of cognitive engagement with the task, people are more likely to respond to simple heuristic rules of thumb such as 'this email is marked as urgent' and 'Oh this is a reputable brand – it must be ok' and are therefore enticed to click on emails that would, if given more thought, appear suspicious.

3. THEORETICAL FOUNDATIONS

In this section, we discuss the theoretical standpoint from which this research was carried out.

We sought to set out from the previously established findings but engage in a more discovery-orientated investigation. Thus we aimed at uncovering the 'unknown unknowns' of this particular space and thereby be better equipped to later select specific models to apply – or develop new models altogether. As such, we deployed qualitative methods, with the aim of gaining insight into the cognitions, work contexts, motivations, normative influences, and everyday practices of staff as they experience phishing attempts. Although our approach was not a full implementation of grounded theory, we deployed many of the techniques prescribed by this method, seeking knowledge from the ground-up and asking questions with varying degrees of specificity in order to probe specific areas of interest.

Qualitative methods are useful for identifying new and undiscovered phenomena, providing deeper insights into user experiences than quantitative measures can provide, may be transferable to populations equivalent to the sample group and can uncover themes that may be later tested with more quantitative approaches [26]. Furthermore, the depth of detail and nuanced, semantic-based responses provide a richer, deeper understanding of the problem-space than offered by higher-*n* quantitative studies with less time devoted to each subject [27].

In this, our research was successful in that it uncovered evidence of both a number of widely reported dynamics and phenomena in the field, as well as promising results that were novel or even contradicted prevailing knowledge in the literature.

4. INSTRUMENT DEVELOPMENT

Starting out from the variables deployed by the theories above, we developed a 38-item questionnaire with questions grouped by topics: Knowledge, Attitudes, History, Practices, Contexts and Identity (see appendix A for the complete instrument). Since we sought to extend our investigation beyond the known, specific constructs of the models in question, we formulated many open questions designed to elicit non-structured, wide-ranging responses. An overview of the literature consulted in the process of developing the questions for each topic is included below.

4.1 Knowledge

Individuals' knowledge of cyber security threats as well as computer literacy and expertise have been proposed as important determinants of protective behaviours. Furthermore, specific variables such as threat awareness and countermeasure awareness have been posited as predictors of IS policy compliance [17]. However, as Stephanou showed [28], while education/training campaigns have measurable impact on staff knowledge of the desired behaviours, they are not necessarily correlated with actual subsequent behaviour suggesting that education is necessary but not sufficient to mitigate victimisation. As such, the questions included in the Knowledge topic were designed to gauge the depths of people's understanding of the domain generally, as well as elicit more emotional and relational perceptions of their role and interactions with others in this context. We wanted to understand how participants thought of and spoke about IS and how it affected them in their everyday work lives. Therefore we developed five questions (Q1,Q1A-D, see Appendix A) and grouped these under the Knowledge topic. These were deliberately broad, open questions, designed to elicit conceptions about security in the most general terms possible and in ways that were most cognitively available and important to participants. Q1-B was designed to elicit conceptions around who were the actors in the IS space – both within the organisation and outside – to try to understand whom participants interacted with and had relationships with in relation to the subject.

4.2 Attitudes

Ever since La Pierre showed that when questions about attitudes are posed broadly they are poor predictors of specific behaviours [29] attitudes have long been known to have a complicated relationship with behaviour. As such we set out to understand how the most commonly implicated variables in IS behavioural models of attitudes were experienced by staff and what kind of situational factors fed into these variables. We were also interested in people's value systems and how ideas about the importance of IS impacted on their intention to comply with mandated security protocols.

Perceptions of vulnerability have been found to be important predictors of people engaging in protective behaviours in a number of models such as the PMT [30]. Perceptions of fear are also central to the threat appraisal process described in this model [11, 31]. Thus we were specifically interested in ideas around vulnerability and fear and formulated questions Q2 to Q4B in a manner that illuminated the contexts in which they are evoked, and categorised these as belonging to the 'Attitudes' Topic.

4.3 History

We were interested in the effects of previous phishing or fraud victimisation on subsequent behaviour and attitudes. Research into related constructs such as 'threat awareness' and 'domain knowledge' has yielded mixed results. Wang et al. [32] showed that prior 'scam knowledge' decreases attention to 'visceral triggers' and increases attention to deceptive elements in fraudulent emails. However Vishwanath et al. [33], deploying a dual process model of information processing, found that domain specific knowledge (a construct including experience and exposure) significantly predicted elaboration likelihood (deeper processing of information according to dual route theories) in only one half of a split-half test population and the relationship was therefore only partially supported. However, in both these cases experience was not directly related to previous victimisation and instead consisted of education and exposure to information without the negative outcomes associated with actual victimisation. On other hand, Böhme et al. [34] found that experience with e-commerce fraud was likely to reduce subsequent on-line purchasing behaviours, but that as a predictor, the effect size was less than 'general concerns' and 'personal concerns'. In a result that may explain some of the variance in findings above, Yu [35] found that victimisation experience significantly affected subsequent fear of cybercrime for cyber bullying, digital piracy and computer viruses – but not for on-line scams, suggesting that post-incident fear perceptions are dependent on the nature of the crime itself.

In light of these seemingly variable findings, we developed a number of questions aimed at understanding participants' real world experience with phishing victimisation and cyber-fraud and how these experiences affected their subsequent and ongoing practices and cognitions around IS (Q5-Q12).

4.4 Practices

We were also interested in staff's experiential relationship with specific known challenges for IS. The questions in this topic were designed to elicit discussion on behaviours around passwords, use of free Wi-Fi, and then, in more detail; email practices (Q13-Q17). The relationship between system usability and restrictive security procedures has been much discussed. For example Post and Kagan [36] showed that increased requirements around the complexity and diversity in user-generated passwords resulted in ever increasing cognitive demands often leading to more risky behaviours (such as writing passwords down). This particular security-usability trade-off is also highly explicable to users and as such, we wanted to understand both their practices and attitudes towards password use as a proxy for more general behaviour around IS.

Since the financial institution involved in our study had a long-standing and significant investment in IS education including an extensive training program, an information portal, phishing drills, and awareness events we were also interested in staff awareness of and response to these engagements (Q18-22).

There has also been discussion about what educational formats are most effective at engaging staff [37] – so we asked about both staff appetite for more learning – and their ideal format for educational materials (Q23).

4.5 Contexts

Behaviour does not occur in a vacuum, and as such we were interested in gaining insight into the environmental and organisational factors that impact on the work practices associated with phishing victimisation. Much of this section was specifically designed to elicit information about staff practices in relation to email as the primary vector for phishing attacks.

Dual route models of information processing suggest that users typically engage in little elaborative information processing when scanning emails, and rely instead of more shallow evaluations based on heuristics such as calls to authority, urgency cues and social proof to make fast decisions about responding [3, 16]. Importantly these attentional-based theories suggest that education is unlikely to be sufficient to curb risky behaviour if the user never engages their implicit knowledge of the subject matter in order to evaluate threats. These attentional models also suggest that a new range of variables – such as workload, attentional resources and task demands (as well as individual differences such as need for cognition [8]) are important determinants of phishing victimisation. Mark et al. [38] showed that some email usage patterns result in users feeling cognitively overloaded and stressed. This, in conjunction with the finding by Vishwanath et al. [33] that the number of emails that users engage with daily (email load) significantly increased the likelihood of falling for phishing attacks, suggest that the sheer volume of emails people respond to provides significant challenges to people's available cognitive resources to evaluate threats. Mark et al. also noted different patterns of behaviour around responding to emails, such as users who process in 'batches' at pre-determined times, to those who check their inbox constantly throughout the day, and those who respond to notifications in real-time. The ramifications of these different patterns of email interaction on phishing responses has yet to be investigated so we formulated Questions 24-29 in order to better understand staff behaviours in this area.

Additionally, two much discussed variables in IS behaviour are those relating to the punitive measures that organisations deploy in response to poor staff security behaviour, specifically perceived certainty of sanctions and perceived sanction severity. These variables are controversial since while prevailing thought within the criminologist domain suggests that increasing these variables leads to more desired behaviours [39] – other studies have shown that they are only weakly predictive [29], particularly when there are avenues to neutralise the effects of their non-compliance [41].

Furthermore, we were also interested in the normative environment in which participants existed and informational and cultural influences on staff attitudes. Dodge et al. [42] amongst many other has shown that staff who have leaders that espouse IS protocols and lead by example are more likely to adopt good practices themselves. Flores et al. [43] showed that both the variables of transformation leadership (where leaders involve subordinates in decision making and driving change) and IS culture were both significant predictors of IS awareness which in turn had significant effects on intrinsic beliefs and then intentions. In terms of social norms and peer influence, Ifinedo [21] deployed Social Bond Theory to show that the four constructs of attachment (to an organisations values), commitment (to an organisations goals), involvement (in an organisations goals) and personal norms were all significant predictors of subjective norms and that subjective norms had a positive effect on attitudes towards IS policy compliance. As such, we asked a number of questions (Q31A-C, Q32A-C) about where people learnt about IS from, and then also about whether they talked about, learnt from or thought of as valued by both peers and managers.

4.6 Identity

In models derived from Ajzen's Theory of Planned Behaviour, as well as later variants such as the PMT, various elements of a person's ideas about their selves, such as locus of control, self-efficacy and response efficacy have been shown to be important predictors of behaviour [44]. Furthermore, motivation has been shown to have a causal relationship with elaborative processing, as expressed in more attention-based models such as the HSM. However, in a finding that poses challenges for motivational-based models, Floyd et al. [23] showed evidence that that self-efficacy was not correlated with elaboration likelihood in evaluating phishing emails whereas level of involvement was.

Therefore, in addition to the ideas about subjective norms which we included in the contexts section above, in this subset of questions (Q33-Q38) we wanted to understand how staff engaged with their work practices in ways that may be affected by such variables such as self-efficacy, response-efficacy, responsibility and locus of control.

5. METHOD

5.1 Participants

18 staff situated in Australia (4) and New Zealand (14) from a major financial services institution, took part in the study (8 female,

10 male). Staff were recruited via emails sent to a cross section of staff. Some organic recruitment also took place as staff began to forward the invitation to colleagues. Staff were offered the opportunity to win double movies passes as recompense for participation, were instructed as to the anonymity protocols involved and informed of the voluntary nature of the experiment via the automated, sign-up web service provided by the bank. Participation consent, and consent for the experimenters making a recording was provided at the beginning of the interview session. The functional roles of the participants are listed in Table 1 below.

Table 1: Functional roles of participants of the study.

Position Category	Number of participants
Customer services, support and sales	4
Technical + Operations	8
Managers	2
Finance and Risk	4

5.2 Apparatus

The 38-item questionnaire had questions grouped by topics: Knowledge, Attitudes, History, Practices, Contexts and Identity (see appendix A for the complete instrument). Many questions were open and designed to elicit extensive, wide-ranging responses. Since we had limited time (30 minutes per interview), insufficient to administer all questions, a randomly selected subset of topics was differentially applied to each participant – with the exception of Knowledge questions – which were applied to all participants. Coverage of topics across participants is detailed in table 2. As many topics were applied to each participant as time permitted. Some participants offered much more detailed, and therefore time-consuming responses than others, leading to an uneven number of topics covered by different participants.

5.3 Iteration

After the first two days of interviews, consisting of 12 participants, an initial analysis of responses was made to determine emergent themes. Based on this analysis, seven additional questions (Q201-207) were developed and then were administered to the remaining participants during the second interview session, referred to as 'round two'. These questions were interleaved with the existing questions according to their topic. The responses arising from these questions are discussed in the results section according to the category that gave rise to each question.

5.4 Procedure

All interviews were carried out remotely with participants 'dialling in' to a conference call from their premises. The interviewers remained at *Data61* premises and were visible via webcam for the first four interviews – but then, after finding that this was of limited utility, for all subsequent interviews only audio was used. Participants only provided audio and were never visible to the interviewers. Throughout the recruitment process, participant anonymity was stressed, and owing to the protocols we deployed, participants were only known to the interviewers by their 'Made-up' ID provided by the bank. This approach seemed to reassure interviewees, and they appeared to speak freely and without evidence of much social desirability bias present in their responses.

5.5 Coding

Interviews were transcribed in full by various authors, with one interview being lost owing to a failure of audio recording equipment. For this participant detailed interviewer notes were used for analysis. All coding was then carried out by the main author, with frequent access to the original recordings for clarification. Coding began with categories suggested by the cognitive variables in related work as detailed in sections 3 and 4. Additional categories were then developed from the data itself as analysis progressed and ones where known variables did not yield fruitful new information were abandoned. For each category identified, the entire body of transcripts was then re-analysed for additional data pertaining to the category uncovered. Further distinctions were made within categories as the data suggested leading in some cases to new questions being developed and deployed in round two of interviews. Eventually a two level taxonomy of findings was established consisting of general categories of responses with sub-themes. Through this process, we achieved saturation, i.e.: a state where little fresh information emerges from subsequent interviews because all the main themes have been uncovered, within our 18 participants [27].

Table 2. Basic demographics and coverage of topics by participant.

	Age bracket	Sex	Years with organisation	Knowledge	Attitudes	History	Practices	Contexts	Identity	Round two questions
P1	45-54	M	6	Yes	P			Yes		
P2	45-54	M	10	Yes	Yes	Yes	P	Yes	Yes	
P3	35-44	M	6	Yes	Yes	P	Yes		Yes	
P4	35-44	F	16	Yes	P		P		P	
P5	35-44	F	8	Yes	Yes		P	Yes		
P6	35-44	M	10	Yes			Yes		Yes	
P7	35-44	F	20	Yes	Yes			P	Yes	
P8	35-44	M	2	Yes			Yes	Yes		
P9	35-44	F	16	Yes	Yes	P	Yes	Yes	Yes	
P10	20-24	M	2	Yes	Yes	P	Yes		P	
P11	25-34	F	7	Yes		P	P			
P12	25-34	M	1	Yes	Yes			Yes		
P13	35-44	F	5	Yes	Yes	Yes	P	Yes	Yes	Yes
P14	45-54	F	1	Yes	Yes	Yes	P	Yes	Yes	Yes
P15	35-44	F	14	Yes	Yes	Yes	Yes	Yes	Yes	Yes
P16	45-54	M	3	Yes	Yes	Yes		Yes		Yes
P17	25-34	M	10	Yes	Yes			Yes	Yes	Yes
P18	>64	M	3	Yes	Yes	Yes		Yes	Yes	Yes

Note: P = Partial coverage of questions for this topic.

6. RESULTS

Results are presented in three categories, organised by the themes emerging from the interviews themselves: information, education and knowledge sharing; experience of email practices; perceptions of threat, vulnerability and responsibility. Note that these categories and themes arise from the data itself and are therefore not directly related to the topics in which the questions were originally grouped.

6.1 Information, education and knowledge sharing

6.1.1 Passive learning is taking place, but active learning needs to be facilitated

We asked several questions designed to establish staff's sources of information about IS and their engagement with and opinions of those sources of information (Q18, 19, 20, 21, 22, 23 and 31A, 31B). We found a wide variety of practices around learning about IS including:

- Learning at specific training events
- As part of their job requirements (noted for staff in more technical and IT related positions).
- The bank's intranet.
- Weekly email bulletins.
- Monthly email bulletins.
- Outside sources of information such as IS websites and third party company security warnings.

Staff generally stated that they received information about IS and viewed this information in a positive light and as a necessary part of their responsibilities within the bank (see more detail on this in section 6.3.5).

But importantly, it should be noted that most of the modes of education staff referenced are passive – and when asked about where staff would go if they had questions about IS issues – there was a high degree of uncertainty.

> *"I don't even know if we've got that kind of stuff on our website." P6*

> *"I said before I don't think we have specific training on cyber crime. There's no specific modules around it…" P7*

> *"I would probably go to* <name of internal corporate intranet - redacted>. *And I would search for security and probably email them or call them and let them know something that had happened, and if they're not the right person then ideally they help you find who the right person is." P14*

6.1.2 Happy to scan an information email for new knowledge

While asking participants about their sources of information on IS we uncovered a recurring pattern of usage of information received via email bulletins. Participants, at all levels of security awareness, expressed positivity about receiving periodic information about IS. When prompted to elaborate on this engagement many responses were characterised by the idea of there not necessarily being much new in the content, but being willing to scan over the information to search for any new pieces of information.

> *"…yeah it's definitely good reminders… It's timely, I don't think it's overwhelming…" P5*

> *"…there's nothing I would read word for word, but I would definitely scan my eyes over it." P13*

> *"I would say it's mostly a repeat. I can see what they're trying to do, and that, the intent of the bank as an organisation needs to insure that all of the staff understands the whole deal. So, the information that comes to where I am is fairly regular." P16*

This finding is an encouraging indicator that staff value periodic information provided by their employer and furthermore have developed nuanced and agentic levels of engagement with these channels to extract information that they see as pertinent to them.

6.1.3 How would people like their information presented? Short, snappy and based on real-life scenarios

Participants experience with information delivery was of particular interest to us so Q23 was specifically crafted to uncover ideas about preferred format of IS information and training. While short videos were mentioned occasionally, most respondents expressed a clear preference for text-based communication and brevity was seen as an important requirement.

Furthermore, a number of respondents all converged on a single underlying theme – the desire for education based on user experiences, outcomes and specific mitigation techniques.

> *"When you're building something around info security training if it's a real life thing that actually happened." P3*

> *"I think something that is a brief short and sharp one or two reduced snippet sort of a thing which says look: 'here is what happens if you did this and here is how you could avoid that'." P8*

> *"But I would really, really ask for something very brief. I feel as if people just LOVE filling up a page with words. And I think bullet points. Can't go past a bullet point." P13*

Participant responses indicated a set of highly specific criteria for information consumption, characterised by ideas of brevity and the desire for information that is tied to their own personal experiences and practices.

6.1.4 Communicating after a bad event – The implications of prior experience

We asked participants about their previous experience with both phishing email victimisation and on-line crime more generally. We were interested in how staff experienced these incidents and what meanings they ascribed to the events and then further how it shaped their ongoing behaviours. (Q6-Q10). In analysing the resultant responses, we noted that responses to Q10 ('Did you tell anyone about it?') varied in what appeared to be a systemic manner that fell into two distinct groups of responses.

> **Group 1:**
>
> Participants reported telling friends and colleagues about a negative IS incident following the fact. In all cases, the stated motivation was to assist others in avoiding the same problem that they experienced. Interestingly in all cases where participants reported broadcasting their negative experience, they also demonstrated high levels of technical awareness and rated themselves as highly competent with computers.

"I certainly did. I spoke to my colleagues, my friends, sort of tried to make sure that people are not getting into that." P8

"Uhh yeah I did tell my colleagues about it, yes. " P17

Group 2:

Participants suggested that they did not want to tell anyone about their experience and specifically thought that it would reflect badly on them. In these cases, participants saw themselves as being less technically adept.

"I may not umm more so if people think 'how dumb she is' <laughs>" P15

"Oh definitely - I was definitely embarrassed. A sense of 'how did I not see that?" P6

6.2 Experience of email practices

6.2.1 Scanning your inbox – where mistakes happen.

When asked about the quantity of emails employees received in a given day, participants volunteered a wide variety of responses ranging from '10' to 'thousands'. However, despite this variety a common theme emerged of the process of quickly 'scanning' their mailbox for important items in order to identify items that were important or time sensitive.

"You tend to - where you might have glanced at an email before and read a few sentences from the subject heading - to know a bit more before you make that judgement - when it's busy and you're stressed - you look at the subject header and you look at the 'to' box and if you're not in there and if you're not called out in the subject as action - you don't look at it." P6

"So if it's someone I'm currently working with I'll read it straight away. If it's like - a general email to a lot of people - then I'll be like 'Well do I have time? Nope - I'll look at it later'." P2

"I would quickly look at who sent it and the content - oh not the content - the subject line and determine whether it's worth me looking at it straight away then I'll flag emails myself to what priority." P5

This finding on its own may not be significant, but when coupled with both the findings about the amount of non-relevant emails (section 6.2.2) and staff's periodic variance in workload (section 6.2.3) – this may be an indication of circumstances when people's cognitive processing of emails is more shallow during busy work periods and are therefore more likely to click harmful links and attachments.

6.2.2 Some people receive high volumes of non-relevant content

Email practices are obviously a primary consideration when investigating staff behaviour in response to phishing attacks. We asked a number of questions designed to elicit staff experiences around practices and contexts when processing incoming emails – both at work and at home (Q15-Q17, Q24-Q29, Q35, Q38). During the initial analysis of session one interviews, we noted a consistent theme emerging where participants would nominate a number of emails that they would receive each day, but then would later modify that amount in respect to how many they thought were actually relevant to them. As a result we added Q207 ('Do you get a lot of emails that aren't really relevant to you? Or are trivial?') to the interviews for participants in session two.

We uncovered that some, but not all, participants talked about having to deal with large numbers of emails that were not particularly relevant to them, or were trivial. These included:

- FYI emails where people were generating a paper trail in order to share responsibility or visibility for a decision or process, but again no action was required of the recipient.
- Spam (non-phishing) emails.
- Magazines and informational emails (presumably via subscription).
- 'Marketing' and promotional emails (presumably un-invited and as a result of submitting user details to an external party).

"Umm yeah a lot of the emails are sort of marketing emails." P17

"Definitely. ... you'd get in any given day where you would skim read it and say 'ok, great, fine, filed'." P6

6.2.3 Periodic variations in workload

Participants were asked several questions focussed around email practices and time pressure at work (Q24 - Q29). After round one of interviews, initial analysis showed that many participants found it difficult to respond to these questions systematically because of the variation in their workload over time. We therefore developed two additional questions (Q201, Q203) that were asked of participants in session two interviews, specifically aimed at exploring this phenomenon.

We found that while a few described their positions as being reasonably stable in terms of workload and time pressure, others indicated a large amount of 'seasonal' variation of these attributes.

"It can get crazy busy, it can get insanely busy and at other times - it can be quite relaxed." P2

"Our days are very umm… no single day is the same." P5

"There are phases when it's very, very busy and you definitely do feel the pressure. But that phase comes once every few weeks. And then it <unintelligible> goes back to normal where it's not so much of a time pressure. ... I think my behaviour changes significantly at that point in time, or during those phases." P8

Furthermore, when we probed deeper and asked participants to expand on their email practices during these different periods – many staff stated that they thought these would vary considerably according to the workload at the time.

"When it's busy and you're stressed - you look at the subject header and you look at the 'to' box and if you're not in there and if you're not called out in the subject as action - you don't look at it." P6

"If it's like - a general email to a lot of people - then I'll be like 'Well do I have time? Nope - I'll look at it later." P2

6.3 Perceptions of threat vulnerability and responsibility

6.3.1 *'At work I feel safe' – Lack of vulnerability*

In order to explore staff feelings around feelings of vulnerability we asked the question: 'Do you feel vulnerable to IS threats?' (Q4A). The majority of responses indicated surprisingly low feelings of vulnerability in response to this question.

> *"Umm at work I feel confident. Umm that our technology team work very hard." P5*

> *"No. Not in the slightest."* (Q4A) *P13*

> *"...probably not so much at work... because I'm pretty sure I feel like we've got good processes in place at work…"* (Q4A) *P4*

The few responses that did indicate some degree of vulnerability were only offered by staff with high levels of cyber security awareness, and were couched in terms of nothing being fully secure.

> *"Well everything is vulnerable - You never know." P12*

> *"So I feel like because I'm aware and I know to speak up about it and double check things, that I am quite safe, myself ... umm however I guess it always plays in the back of your mind." P6*

6.3.2 *Information security in home contexts – Far more vulnerable, but less to lose*

While we noted a high degree of confidence in the bank's security apparatus generally to protect them from the worst of information attacks, (see section 6.3.1), there seemed to be an acute consciousness amongst staff interviewed that these defence mechanisms were not available at home or on their personal devices. Thus for interview session two we added Q202: 'So what is the difference between thinking about, or IS practices at home compared to at work?' Responses to this question reinforced the finding that feelings of vulnerability were higher at home than at work.

> *"Whereas at home - you're that person that is susceptible to all those things - and those safety measures that the organisations put in place so therefore you think that much more about it. Or you SHOULD think that much more about it." P6*

> *"And personally - umm - outside of work umm <laughs> - not so protected!" P5*

> *"I might be even more conscious because I know that if anything goes wrong I'm going to have to sort it out - whereas at work if it goes wrong at least we have support networks to help us." P9*

Additionally many participants reported differences in the sense of ownership/responsibility of the problem-space compared with at work. This was particularly true for employees who take on a lead role in managing IT systems for their household.

> *"At home you are tech security - well I am. <laughs> Whereas at work I'm not." P3*

> *"I'm forever telling my wife of the latest scam that's happened." P6*

But counter-intuitively, participants often reported more permissive and less stringent IS behaviour in the home and specifically talked about this in the context of the consequences being less important. This finding is mysterious and requires further investigation – specifically operationalising constructs around locus of control, vulnerability and threat severity.

> *"…but at home there's more risk because I probably don't have as strong a firewall." P7*

> *"When I'm at home, I'm a bit more loose with my emails but I don't click on links." P10*

6.3.3 *Got scammed? Money was returned so no real loss! Lack of vulnerability*

After investigating feelings of vulnerability, we extended our research into the area of perceived consequences of IS breaches. After identifying those participants who had experienced an episode of cyber fraud, we probed extensively into their experiences and reactions (Q5-Q9). We found that in all cases, respondents reported that the fraudulent transactions affecting their accounts were reversed by the bank. In some cases this happened quite quickly, while in one case only a partial reimbursement took place. All staff had banking accounts with the financial institution in question. When discussing these events, respondents were highly positive about their bank's response and offered high estimations of the bank's processes in these cases.

> *"But* <bank name redacted> *were brilliant. Seriously. Within 24 hours I think they had replaced the limit and <unintelligible> take care of it. So for me -whatever happens after that, doesn't really matter." P13*

> *Interviewer: "In your case there was no consequence because it was stopped pretty much immediately, is that right?"*

> *Participant: "Yep." P18*

This can be seen to equating to the much discussed variable of threat severity which has also been discussed extensively as contributing (when high) to protective behaviours.

6.3.4 *Security failures equated with loss of trust in the bank*

In order to try to understand ideas of loss and the perceptions of consequences of poor IS behaviours we asked Q1d ('Why is information security important?') and then also Q12 ('What is the worst thing that could happen as a result of a phishing attack on you?'). As well as a host of responses detailing specific worst case scenarios centred around data loss and fraud, we noticed a repeated theme amongst many of the participants who volunteered that one of the worst consequences of successful attacks would be the reputation of the bank itself. It would appear that staff are highly cognisant of the wider implications data breaches and place a consistently high value on the reputation of the bank.

> *"…because at the end of the day - it is going to affect the um - what is the word I'm looking for - the name of my employer. So <bank name redacted> at the end of the day will be affected and we don't want it to be named and shamed in any way. So the reputation is at risk." P15*

> *"We're a bank, banks are built on trust, if we don't have the trust of our customers, we're out of business." P1*

"And the potential is, if we do it wrong, really badly wrong - and we lose that money - it's not a good thing. And I think primarily trust. Customer trust in us." P3

"Because as a bank, we have a very high trust mandate…" P16

6.3.5 Responsibility for security – and identifying with the bank

Since a higher internal locus of control has been shown to be a necessary but not sufficient antecedent of engaging protective behaviours we operationalised this concept in several questions designed to elicit staff attitudes towards who was responsible for preventing IS attacks (Q1b, Q1d, Q3). During the initial analysis and iteration process after interview session one, we noted that many people talked about this issue quite specifically in relation to their perceived identity – as an employee of a bank, so we added Q204 ('So you work for a bank – does that bring with it any special responsibilities and roles?') for interview session two.

We noted that staff universally offered two primary observations about their perceived responsibility for IS. Firstly – and almost always offered as a response to Q1b – people volunteered that they, as an individual, were the primary actor and determinant in this space.

"Me as an individual I am primarily responsible for my own security…" P14

"Well I think I'm the primary." P13

"It's yourself and anybody who's responsible for public and private networks, and the ownership of those." P18

"I think it starts with you as a person. But I think everyone is involved…" P11

This assertion of individual responsibility was then almost always followed by a secondary consideration – that of a shared responsibility with others – primarily the bank – but often institutions generally and the collective ideas of the staff at large – all seen as powerful outside forces in the equation.

"So, cybersecurity as a holistic level really comes back to the education of everyone." P16

"…everybody's! <laughs> Everybody who is involved in the network and anybody who's responsible for their own approaches <unintelligible> and use of the system." P18

"Everyone should be aware of what's happening…." P2

"Protect my details as much as I can possible, I absolutely would take full responsibility for that. But at the same time I'm happy to lean on the bank when things do go pear shaped." P13

7. DISCUSSION

Analysis of participant responses revealed a number of novel observations as well as confirmed some findings well established in the literature. Here we discuss the wider implications of these findings on future work in this domain as well as possible real-world applications to combat cyber-crime.

7.1 Information, education and knowledge sharing.

Since knowledge of cyber threats has been shown as a necessary but not sufficient pre-requisite for users carrying out appropriate protective measures [40], the insights established when asking about how users receive and participate in information sharing are encouraging. In section 6.1.1 we learnt about how participants' learnt about IS and engaged with a wide variety of sources of information. However, much of this information seemed to not comply with basic instructional design principles such as those discussed by Kumaraguru et al. [15]. Furthermore when asked about where users would go if they had questions, there was much uncertainty, suggesting that, in this particular context, more active modes of information acquisition should be further facilitated. In particular, this reflects a lack of the basic instructional design principles iterated by Kumaraguru et al. where training was most likely to be effective when offered at the right time – i.e.: when participants were interested in learning or those moments when the information is particularly salient. This distinction between passive and active learning raises interesting questions about what additional protocols can be developed to meet staff needs and how active learning behaviours varies across the population in question.

In section 6.1.2, we showed that the participants we spoke to were not fatigued by regularly scheduled information broadcasts from the bank and that these established communications conduits were perceived as useful. More specifically, a behaviour pattern that repeatedly presented itself whereby users would scan over periodic IS related bulletins, assess what was relevant to them and then engage with new material that they deemed relevant to them, indicating a nuanced and agentic engagement with these sources of information. This process is also interesting in relation to our finding in 6.1.3 where the desired modes of information presentation (brief and based on personalised and highly situated stories) were both immediately cognitively available to users as desirable and quite specific, again reflecting Kumaraguru's et al. [15] instructional design principles of personalisation, contiguity and story based agent environment as well as the findings by Harbach et al. [45] on the effectiveness of information personalisation. This reinforces the need for increasing the bandwidth of existing channels of information to staff by adopting principles of brevity and presenting narrative based information based on specific user personas. Furthermore, this supports an emerging trend by industry where education efforts are personalised for different categories of user in order to increase relevance and engagement.

When asking about information sharing with peers generally (section 6.1.4), no consistent pattern of responses occurred, however when probing into experiences of cyber-crime victimisation, an interesting theme emerged. Staff with a self-image of being highly computer–literate and technically 'savvy' seemed more likely to communicate to others about their negative experiences, than those with less certainty about their technical skills. For those with low technical self-efficacy, cognitions around victimisation were more centred around the risk of appearing foolish or careless if they talked about their experience. Since an oft stated goal of IS culture is to increase normative information sharing [46] - from an organisational standpoint this would imply that reaching out to those with high levels of confidence in their technical abilities may facilitate the development of localised 'champions of change' throughout the organisation. This approach would extend the work by Sauvik Das et al. in the area of social proof, where normative influence was shown to significantly increase uptake of additional FB security features. Furthermore this approach specifically suggests avenues to overcome the significant challenge for deploying on social proof dynamics where engaging in protective technologies is not easily visible to others.

7.2 Experience of email practices

While the HCI literature has grappled extensively with the phenomena of email processing and problems such as overload [38] and interaction patterns such as task switching, interruption lag and resumption lag [47], leveraging these findings in order to mitigate phishing threats has remained elusive. However, more recently, models emphasizing the attentional aspects of phishing victimization have appeared promising. The Heuristic Systematic Model (HSM) with its focus on competing (and interacting) shallow and deep information processing mechanisms has been shown to predict some degree of phishing victimization. As such our discovery of a number of real-world phenomena present in staff email usage that are likely to impact on the application of models such as the HSM to the problem may prove useful in future efforts.

We found evidence of shallow processing of incoming emails when people talked about 'scanning' their inbox for emails that needed an immediate response (section 6.2.1). This reflects Neustadter's behavioural findings of the email triage process [48], such as the common tactic of attempting to remove trivial emails in order to make it easier to find more important ones. However, this behavioural analysis needs to be extended by a deeper understanding of the mental processes involved in order to effectively mitigate phishing victimisation. Models based on attentional theories, such as the HSM offer good utility here. An example can be found in Xu's [8] exposition of the likelihood of elaborative processing in email processing according to personality traits such as need for cognition and contextual variables such as recipient expertise and recipient involvement (a motivational factor). Based on a similar dual process model, Vishwanath et al. [33] showed that most email is processed peripherally and that SE was an important factor in users engaging in elaborative processing. Furthermore, in findings that extend the exposition by Neustaeder et al. on the email triage process, Floyd [23] showed that that calls to scarcity or urgency cues in phishing emails resulted in higher levels of user response owing to the dominance of peripheral/shallow processing strategies.

This understanding of engagement with incoming emails is likely to be complicated by our finding of substantial variations in workload over time (see section 6.2.3). Participants asserted that the way they process their inbox is qualitatively different depending on how heavy their workload is at the time – possibly explaining some of the variation in the effectiveness of other known predictors of systemic information processing such as desired confidence and motivation. This may also go some way to explaining a much discussed phenomena in the security services industry around phishing emulations where specific emails that have been 'benchmarked' according to their effectiveness, for use as calibration tools across organisations, nonetheless exhibit a wide degree of variance in victimisation rates.

In addition, we discovered that participants varied substantially in their reported numbers of non-relevant emails they received each day (see section 6.2.2). This reinforces the relevance of the assertion by Parson et al. [49] that the categories of emails that user needs to process may have a profound effect on the mental processes involved and leads to a great deal of uncertainty in terms of experimental design. It further complicates Neustaedter's et al. [48] taxonomy of low, medium and high volume users and suggests another variable that may need to be accounted for to explain processing approaches. At the very least, the proportion of non-relevant/trivial emails is likely to effect the mental efforts devoted to evaluating each email – and if high levels of non-relevant emails results in shallow processing, may result in increased victimisation as users devote less elaborative processing to evaluating the characteristics of each email. This variable should be operationalised and tested in further attentional-based experiments into phishing victimisation.

7.3 Perceptions of threat, vulnerability and responsibility

Vulnerability and threat severity.

We found participants consistently talked about feeling 'safe' and 'protected' within the information infrastructure of the bank – and tied these feelings firmly to the perceived emphasis and obvious presence of IS protocols, practices and information in their workplace. An anecdote related to us by a bank security worker involved a staff member who noticed a suspicious email they received in their personal email account at home, and had immediately forwarded it to their professional email address in order to open it at work – rationalising this as the safest thing to do since the security environment at work was 'safer' than that which they had access to at home.

These low feelings of vulnerability owing to the perceived presence and visibility of IS practices suggests itself as an important finding since it ties in with the literature around risk homeostasis. This theory suggests that in situations of risk, where controls are implemented to mitigate the risk or the severity of the outcomes, people often either decrease their protective behaviours, or increase risky behaviours in order to subconsciously return to the same level of risk as before the mitigation was put in place [50]. This effect has been seen in examples such as anti-locking brakes, where drivers, once aware of the effect of the new braking system on stopping distance, modified their behaviour to drive closer to cars in front of them – returning the risk to subjectively the same levels as before the application of the protective technology - the anti-locking brakes [51]. An emerging design response to this dynamic has been increasing the subjective feelings of risk and vulnerability in order to encourage users to engage in protective behaviours. This has been implemented in domains as far flung as traffic calming designs to aviation systems [52]. This has immediately actionable implications for staff education in that according to all of the models that include the concept of vulnerability, emphasising this variable in educational efforts is likely to increase protective behaviours deployed by staff. On the other hand, whether this mitigation approach is palatable to organisations' internal communication values is debatable.

When asked about the difference between practices at home and at work (section 6.3.2) we found indications that perceptions of vulnerability were higher at home than at work while threat severity was lower at home than at work. These assertions were specifically linked to the perception that at home breaches would be centred around personal loss, but at work would also potentially damage the bank as well.

This understanding of threat severity related to the bank itself was also evident in responses to questions in section 6.3.4 where staff placed a heavy emphasis on the consequences to the reputation and trust of the bank by its clients should it experience a major security breach.

Again, models based in the cognitive literature have been shown to explicate such findings as well as offer avenues for mitigation. Boss's deployment of PMT in a study of virus alert warning messages, showed that perceived threat severity predicts fear, which in turn increases (in the described study by double) the perception of the necessity of taking protective measures. On the other hand, this particular variable is contentious as a predictor since Hanus et al. [17] found that that it was not a signification predictor of security behaviour.

This variance in evidence may be explained by a central dynamic of both the PMT and HBT where they predict that threat severity will only increase protective measures when SE is high. I.e.: regardless of how motivated people are to protect themselves, they will not do so if they believe that are not capable of carrying out the necessary actions to protect themselves.

We found that participants often exhibited very little fear about monetary loss in response to cyber attacks. This was evident in a large number of responses where participants detailed being the victim of cyber-fraud, but with the final outcome of their money being replaced by the bank – sometimes very quickly. This notion is also supported by repeated assertions about faith in the bank to replace lost funds should something go awry.

Furthermore, several participants, when discussing their own fraud victimisation, repeatedly used the term reduced 'limit' to describe the outcome of the event and did not seem to perceive that attack as actually involving any monetary loss. This would suggest that people see losses charged to a credit card as qualitatively different and of far less consequence than that of losses to a savings type account.

These two phenomena together may help explain Yu's [35] finding that victimisation experience significantly increased subsequent fear of cybercrime for cyber bullying, digital piracy and computer viruses – but not for on-line scams.

Locus of Control and responsibility

Staff perceptions around responsibility are particularly interesting where while the primary assertion of responsibility was expressed as lying with the individual, it was then immediately qualified by equally strong assertions of a more collective and dispersed responsibility. While the existent HCI literature does not seem to have engaged with the variable of Locus of Control (LOC) directly, it has been shown, within cognitive studies, to be an important predictor of people engaging in protective activities [53].

However LOC is a complicated construct. Walston showed evidence that it is not a unidimensional continuum but rather two independent constructs [54], and since then a number of researchers have attempted to tease out the proposed multi-dimensional space at the nexus of what has been variously called: Self Efficacy, Perceived behavioural control, Locus of Control and Locus of Responsibility [55].

After Rotter's [56] original formulation of the Locus of Control, Levenson [57] extended the model by proposing three subscales: internality, control by powerful others and control by chance. These variables suggest themselves as being particularly apt to this context since staff seem to put much stock in the presence of existing security systems and protocols – implying awareness of the presence of powerful others.

This is however complicated by the fact that in Levenson's [57] formulation, the presence of powerful others are more likely to be considered agentic in the outcome in question, whereas in our context the presence of powerful others, in the form of effective security systems of the bank, suggests an attribution of less likelihood of the reinforcement – i.e.: falling victim to phishing.

There may also be reason to entertain Terpstra's [58] distinction between moral and actionable responsibility. We saw responses where participants discriminated between taking personal responsibility for engaging protective behaviours, but then relying on the bank to provide technical and material assistance. This suggests that there is some perceived distinction made between the roles of the individual and the bank that may correspond to individual actions being seen as a moral responsibility, but organisational responses as more agentic and actionable.

A further argument for the importance of outcome attribution is presented by Jeuring et al. [59], who deployed an additional variable of Locus of Responsibility (LoR) and showed that internal LoR is associated with higher engagement of coping strategies, but only if it is also accompanied by a perception that the person has the necessary resources to mitigate the risk, ie: Self-efficacy.

8. LIMITATIONS

This research, being qualitative, resulted in a number of findings that should be considered not as generalised facts, but rather understandings of processes, in a particular context of a particular group of people in a particular industry. While the nature of this knowledge is richer and deeper than that typically resulting from more quantitative approaches, questions of generalisability remain to be addressed by further more quantitative and larger-*n* work as discussed below [60].

Furthermore, since our study took place within a specific socio-technical system, i.e.: a large bank in Australia and New Zealand, it remains to be seen as to how inter-organisational and inter-cultural differences may affect these findings. Specifically cognitions around punishment for maleficence and IS policies generally are likely to vary from institution to institution and ideas around sharing information and identity may vary across cultures.

9. CONCLUSIONS

Our study resulted in a number of findings that suggest both avenues for future research and intriguing hypotheses to test. We present here a summary of our work related back to the original research questions provided at the outset of this paper.

What cognitive variables may be implicated in staff's behaviour in relation to phishing emails?

We found that self-efficacy may well be a strong determinant of staff sharing stories of negative experiences. This is owing to those staff with a self-image of being less technically literate being embarrassed to admit victimisation while those who saw themselves as technically competent felt motivated to share their stories to prevent victimisation of their peers. In terms of perceptions of threat and vulnerability, we found a noteworthy lack of perceived vulnerability when within the bank's IT systems that were associated with impressions of confidence in the bank's visible and highly estimated security protocols. Low perceptions of vulnerability within bank networks were often accompanied by stories of falling victim to identity theft but where financial loss was quickly mitigated by the bank – leading to a postulated low threat severity attribution specifically for financial victimisation.

How do staff experience information security within the organisation and how does this differ from their perceptions at home?

We found that mitigating IS risks was perceived as a shared responsibility between the individual and the wider bank security systems. Staff conceptions around security breaches were heavily centred around cognitions of subsequent loss of trust in the bank by the public and was seen as an important and central issue for employees. In contrast to the above finding about low feelings of vulnerability within the banks networks, we found different perceptions around on-line experiences at home where participants felt more vulnerable, but where a wide range of perceptions around threat severity was found.

What environmental and organisational factors affect staff behaviour in relation to phishing attacks and information security more generally?

We found that in relation to education, the existing informational channels seemed to be functioning and well received. However there was opportunity to capitalise on staff self-motivation by providing more avenues for active learning and that participants expressed a clear preference for information presented in brief stories centred around personalised experiences and work contexts. In relation to email practices, we found that some staff receive far more emails than others and that there appears to be much variance in the proportion of non-relevant, or trivial emails that staff receive on a day to day basis which has implications for attentional models of information processing in relation to phishing victimisation. Workload was also found to vary significantly over time for some staff, and that this was associated with perceived differences in practices around scanning and responding to incoming messages.

10. FUTURE WORK

The research presented here describes a number of novel observations pertaining to banking staff cognitions around and experiences of IS. While these findings suggest further investigation, we mention three of the more promising avenues for further research below.

Our finding that staff with a self-image of computer competence and being technically 'savvy' are more likely to communicate to others about their negative experiences (section 6.1.4) should be investigated further. Understanding what factors preclude people from discussing and sharing information about phishing victimisation holds promise for creating organisational cultures with increased normative influence on staff about the correct protective behaviours to deploy.

Our finding in relation to variance in the number of non-relevant emails staff encounter in their inbox may have important implications for attentional based and dual process theories of phishing victimisation. This variable should be deployed in future work employing theories such as the Heuristic Systematic Model to predict elaborative processing of incoming emails.

Perhaps most interestingly, our findings of low feelings of vulnerability associated with visible organisational security protocols suggests an important avenue for staff education efforts. Manipulating vulnerability in messaging and then validating via behavioural responses may increase protective measures as predicted by risk homeostasis theory.

On a more general note, we suggest that modes of investigation that consider and deploy cognitive variables are likely to be of considerable benefit to the HCI and CSCW communities. Specifically attacks based around social engineering require an understanding of the mental processes that result in victimisation, and in the context of phishing, the factors that lead to elaborative processing; i.e.: users actually deploying the knowledge that they have to evaluate threats. We argue that models based on the fundamental mental constructs that drive behaviour are likely to be increasingly useful in combatting the ever-increasing sophistication of on-line threats and hold promise to transform users from a system weakness to an active line of defence.

11. ACKNOWLEDGMENTS

The authors would like to thank the information security staff at the institution discussed for their support, and the participants for their fearless contributions.

12. REFERENCES

[1] "IBM Security Services 2014 Cyber Security Intelligence Index," 12-May-2015. [Online]. Available: http://www.ibm.com/developerworks/library/se-cyberindex2014/index.html. [Accessed: 06-Mar-2017].

[2] W. Melicher *et al.*, "Usability and security of text passwords on mobile devices," presented at the Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016, pp. 527–539.

[3] B. Ur, J. Bees, S. M. Segreti, L. Bauer, N. Christin, and L. F. Cranor, "Do Users' Perceptions of Password Security Match Reality?," presented at the Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016, pp. 3748–3760.

[4] A. Acquisti and J. Grossklags, "Privacy and rationality in individual decision making," *IEEE Secur. Priv.*, vol. 3, no. 1, pp. 26–33, 2005.

[5] S. Egelman, "My profile is my password, verify me!: the privacy/convenience tradeoff of facebook connect," presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2013, pp. 2369–2378.

[6] D. Krawczyk, J. Bartlett, M. Kantarcioglu, K. Hamlen, and B. Thuraisingham, "Measuring expertise and bias in cyber security using cognitive and neuroscience approaches," in *2013 IEEE International Conference on Intelligence and Security Informatics*, 2013, pp. 364–367.

[7] M. Workman, "Wisecrackers: A theory-grounded investigation of phishing and pretext social engineering threats to information security," *J. Am. Soc. Inf. Sci. Technol.*, vol. 59, no. 4, pp. 662–674, Feb. 2008.

[8] Z. Xu and W. Zhang, "Victimized by Phishing: A Heuristic-Systematic Perspective," *J. Internet Bank. Commer.*, vol. 17, no. 3, pp. 1–16, Jan. 1970.

[9] X. (Robert) Luo, W. Zhang, S. Burd, and A. Seazzu, "Investigating phishing victimization with the Heuristic–Systematic Model: A theoretical framework and an exploration," *Comput. Secur.*, vol. 38, pp. 28–38, Oct. 2013.
[10] Y. Sawaya, M. Sharif, N. Christin, A. Kubota, A. Nakarai, and A. Yamada, "Self-Confidence Trumps Knowledge: A Cross-Cultural Study of Security Behavior," presented at the Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 2017, pp. 2202–2214.
[11] S. R. Boss, D. F. Galletta, P. B. Lowry, G. D. Moody, and P. Polak, "What Do Users Have to Fear? Using Fear Appeals to Engender Threats and Fear that Motivate Protective Security Behaviors," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 2607190, Dec. 2015.
[12] R. Dhamija, J. D. Tygar, and M. Hearst, "Why phishing works," presented at the Proceedings of the SIGCHI conference on Human Factors in computing systems, 2006, pp. 581–590.
[13] S. Preibusch, K. Krol, and A. R. Beresford, "The privacy economics of voluntary over-disclosure in Web forms," in *The Economics of Information Security and Privacy*, Springer, 2013, pp. 183–209.
[14] N. Christin, S. Egelman, T. Vidas, and J. Grossklags, "It's all about the Benjamins: An empirical study on incentivizing users to ignore security advice," presented at the International Conference on Financial Cryptography and Data Security, 2011, pp. 16–30.
[15] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong, "Teaching Johnny not to fall for phish," *ACM Trans. Internet Technol. TOIT*, vol. 10, no. 2, p. 7, 2010.
[16] J. Nielsen, "User education is not the answer to security problems," *Alertbox Oct.*, 2004.
[17] B. Hanus and Y. "Andy" Wu, "Impact of Users' Security Awareness on Desktop Security Behavior: A Protection Motivation Theory Perspective," *Inf. Syst. Manag.*, vol. 33, no. 1, pp. 2–16, Jan. 2016.
[18] M. Fishbein and I. Ajzen, *Predicting and Changing Behavior: The Reasoned Action Approach*. Taylor & Francis, 2011.
[19] I. Ajzen, "From Intentions to Actions: A Theory of Planned Behavior," in *Action Control*, P. D. J. Kuhl and D. J. Beckmann, Eds. Springer Berlin Heidelberg, 1985, pp. 11–39.
[20] S. K. Steginga and S. Occhipinti, "The Application of the Heuristic-Systematic Processing Model to Treatment Decision Making about Prostate Cancer," *Med. Decis. Making*, vol. 24, no. 6, pp. 573–583, Nov. 2004.
[21] P. Ifinedo, "Information systems security policy compliance: An empirical study of the effects of socialisation, influence, and cognition," *Inf. Manage.*, vol. 51, no. 1, pp. 69–79, Jan. 2014.
[22] A. Vance, M. Siponen, and S. Pahnila, "Motivating IS security compliance: Insights from Habit and Protection Motivation Theory," *Inf. Manage.*, vol. 49, no. 3–4, pp. 190–198, May 2012.
[23] D. L. Floyd, S. Prentice-Dunn, and R. W. Rogers, "A Meta-Analysis of Research on Protection Motivation Theory," *J. Appl. Soc. Psychol.*, vol. 30, no. 2, pp. 407–429, Feb. 2000.
[24] J. M. Davis and B. M. Tuttle, "A heuristic–systematic model of end-user information processing when encountering IS exceptions," *Inf. Manage.*, vol. 50, no. 2, pp. 125–133, 2013.
[25] M. Butavicius, K. Parsons, M. Pattinson, and A. McCormac, "Breaching the Human Firewall: Social engineering in Phishing and Spear-Phishing Emails," *ArXiv160600887 Cs*, May 2016.
[26] E. M. Trauth, *Qualitative Research in IS: Issues and Trends*. IGI Global, 2001.
[27] I. Dey, *Qualitative Data Analysis: A User Friendly Guide for Social Scientists*. Routledge, 2003.
[28] A. Stephanou, "The impact of information security awareness training on information security behaviour," Thesis, 2009.
[29] R. T. LaPiere, "Attitudes vs. Actions," *Soc. Forces*, vol. 13, no. 2, pp. 230–237, 1934.
[30] H. Boer and E. Sydel, "Protection Motivation Theory," *Prot. Motiv. Theory*, pp. 95–120, 1996.
[31] S. Milne, S. Orbell, and P. Sheeran, "Combining motivational and volitional interventions to promote exercise participation: Protection motivation theory and implementation intentions," *Br. J. Health Psychol.*, vol. 7, no. 2, pp. 163–184, May 2002.
[32] J. Wang, T. Herath, R. Chen, A. Vishwanath, and H. R. Rao, "Research Article Phishing Susceptibility: An Investigation Into the Processing of a Targeted Spear Phishing Email," *IEEE Trans. Prof. Commun.*, vol. 55, no. 4, pp. 345–362, Dec. 2012.
[33] A. Vishwanath, T. Herath, R. Chen, J. Wang, and H. R. Rao, "Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model," *Decis. Support Syst.*, vol. 51, no. 3, pp. 576–586, Jun. 2011.
[34] R. Böhme and T. Moore, "How do consumers react to cybercrime?," in *2012 eCrime Researchers Summit*, 2012, pp. 1–12.
[35] S. Yu, "Fear of Cyber Crime among College Students in the United States: An Exploratory Study," *Int. J. Cyber Criminol.*, vol. 8, no. 1, p. 36, Jan. 2014.
[36] G. V. Post and A. Kagan, "Evaluating information security tradeoffs: Restricting access can interfere with user tasks," *Comput. Secur.*, vol. 26, no. 3, pp. 229–237, May 2007.
[37] P. Kumaraguru, Y. Rhee, A. Acquisti, L. F. Cranor, J. Hong, and E. 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*, New York, NY, USA, 2007, pp. 905–914.
[38] G. Mark, S. T. Iqbal, M. Czerwinski, P. Johns, A. Sano, and Y. Lutchyn, "Email Duration, Batching and Self-interruption: Patterns of Email Use on Productivity and Stress," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, 2016, pp. 1717–1728.
[39] Q. Hu, Z. Xu, T. Dinev, and H. Ling, "Does Deterrence Work in Reducing Information Security Policy Abuse by Employees?," *Commun ACM*, vol. 54, no. 6, pp. 54–60, Jun. 2011.
[40] J. S. Downs, M. Holbrook, and L. F. Cranor, "Behavioral Response to Phishing Risk," in *Proceedings of the Anti-phishing Working Groups 2Nd Annual eCrime Researchers Summit*, New York, NY, USA, 2007, pp. 37–44.
[41] M. Siponen and A. Vance, "Neutralization: New Insights into the Problem of Employee Information Systems Security Policy Violations," *MIS Q.*, vol. 34, no. 3, pp. 487–502, 2010.
[42] R. Dodge, K. Coronges, and E. Rovira, "Empirical Benefits of Training to Phishing Susceptibility," in *Information Security and Privacy Research*, 2012, pp. 457–464.

[43] W. Rocha Flores and M. Ekstedt, “Shaping intention to resist social engineering through transformational leadership, information security culture and awareness,” *Comput. Secur.*, vol. 59, pp. 26–44, Jun. 2016.
[44] I. Ajzen, “The theory of planned behaviour: Reactions and reflections,” *Psychol. Health*, vol. 26, no. 9, pp. 1113–1127, Sep. 2011.
[45] M. Harbach, M. Hettig, S. Weber, and M. Smith, “Using personal examples to improve risk communication for security & privacy decisions,” presented at the Proceedings of the 32nd annual ACM conference on Human factors in computing systems, 2014, pp. 2647–2656.
[46] S. Das, A. D. Kramer, L. A. Dabbish, and J. I. Hong, “Increasing security sensitivity with social proof: A large-scale experimental confirmation,” presented at the Proceedings of the 2014 ACM SIGSAC conference on computer and communications security, 2014, pp. 739–749.
[47] A. Gupta, R. Sharda, and R. A. Greve, “You’ve got email! Does it really matter to process emails now or later?,” *Inf. Syst. Front.*, vol. 13, no. 5, pp. 637–653, 2011.
[48] C. Neustaedter, A. Brush, and M. A. Smith, “Beyond from and received: Exploring the dynamics of email triage,” presented at the CHI’05 extended abstracts on Human factors in computing systems, 2005, pp. 1977–1980.
[49] K. Parsons, A. McCormac, M. Pattinson, M. Butavicius, and C. Jerram, “The design of phishing studies: Challenges for researchers,” *Comput. Secur.*, vol. 52, pp. 194–206, Jul. 2015.
[50] G. J. S. Wilde, “Risk homeostasis theory: an overview,” *Inj. Prev.*, vol. 4, no. 2, pp. 89–91, Jun. 1998.
[51] C. M. Farmer, A. K. Lund, R. E. Trempel, and E. R. Braver, “Fatal crashes of passenger vehicles before and after adding antilock braking systems,” *Accid. Anal. Prev.*, vol. 29, no. 6, pp. 745–757, Nov. 1997.
[52] K. Malnaca, “Risk Homeostasis Theory in Traffic Safety,” *21st ICTCT Workshop*, vol. Session IV: A theoretical approach, pp. 1–7, 1990.
[53] N. Lalwani and T. S. Duval, “The Moderating Effects of Cognitive Appraisal Processes on Self-Attribution of Responsibility,” *J. Appl. Soc. Psychol.*, vol. 30, no. 11, pp. 2233–2245, 2000.
[54] K. A. Wallston, “Assessment of control in health-care settings,” in *Stress, personal control and health*, A. Steptoe and A. Appels, Eds. Oxford, England: John Wiley & Sons, 1989, pp. 85–105.
[55] K. Wallston, “Control Beliefs: Health Perspectives,” *MyScienceWork*.
[56] J. B. Rotter, *Social learning and clinical psychology*, vol. ix. Englewood Cliffs, NJ, US: Prentice-Hall, Inc, 1954.
[57] H. Levenson, “Reliability and Validity of the I,P, and C Scales - A Multidimensional View of Locus of Control.,” Aug. 1973.
[58] T. Terpstra, “Flood Preparedness Thoughts, feelings and intentions of the Dutch public,” University of Twente, 2010.
[59] J. Jeuring and S. Becken, “Tourists and severe weather – An exploration of the role of ‘Locus of Responsibility’ in protective behaviour decisions,” *Tour. Manag.*, vol. 37, pp. 193–202, Aug. 2013.
[60] E. Albrechtsen, “A qualitative study of users’ view on information security,” *Comput. Secur.*, vol. 26, no. 4, pp. 276–289, Jun. 2007.

APPENDIX

A. COMPLETE INSTRUMENT

Topic: Knowledge

Q1 What do you know about cyber security?

Q1-A What do you think it is?

Q1-B Who is involved?

Q1-C Is it important?

Q1-D Why is it important?

Topic: Attitudes

Q2 What is your role at <bank name redacted>?

Q3 How does Cyber Security affect you?

Q4 How do you feel about Cyber Security?

Q4A Do you feel vulnerable to cyber security threats?

Q4B Do you feel fear about Cyber Security?

Topic: History

Q5 What’s your experience with Cyber Security historically?

Q5A What kind of stories have you heard?

Q6 Have you ever clicked on something dodgy? What happened?

Q7 Have you clicked on a phishing email? What happened?

Q7A How severe was it?

Q8 Did that make you change your behaviour?

Q9 How did that make you feel?

Q10 Did you tell anyone about it?

Q11 Has it happened again since?

Q12 What do you think is the worst thing that could happen as a result of a phishing attack on you?

Topic: Practices

Q13 How do you manage passwords?

Q14 Do you connect to free Wi-Fi?

Q15 What makes you suspicious of an email? Discuss.

Q16 How do you deal with emails you are suspicious of?

Q17 Is it getting hard to tell what is suspicious?

Q18 Where do you learn or hear about this stuff?

Q19 Whom do you trust for advice or information on Cyber Security?

Q20 Do you follow their advice?

Q21 Do you think there is enough training/information provided at your work?

Q22 Would you like to learn more?

Q23 How would like this training/information to be provided? (prompt: video, podcasts, intranet pages, workshops, induction?)

Q24 Roughly how many emails would you receive in an average work day?

Topic: Contexts

Q25 Roughly how many emails would you send in an average work day?

Q26 How do you feel about your email practices (is it too much, stressful)?

Q27 In your email practice, do you tend to; Check/Notifications/Batch

Q28 How busy do you feel at work? Do you feel you have enough time in your day to devote to each task you need to do?

Q29 How stressed do you feel at work on an average day?

Q30 Are there any consequences at <bank named redacted> for poor security behaviour?

Q31 Colleagues:

Q31A Do you talk about cyber-security issues?

Q31B Have you learnt things from them?

Q31C Do they seem to care about cyber security?

Q32 Bosses:

Q32A Do they talk about cyber-security issues?

Q32B Have you learnt things from them?

Q32C Do they seem to care about cyber security?

Topic: Identity

Q33 Do you see yourself as being good with computers?

Q34 Are you confident with your use of the internet?

Q35 Do you think you can recognise dodgy emails?

Q36 Do you teach or tell other people about Cyber security matters?

Q37 Whose responsibility is it to prevent Cyber Security attacks?

Q38 Is it important to you to be able to recognise dodgy emails?

Topic: Iterated – Round two interviews only

Q201 So does the way you scan your inbox change according to how busy you are? And if so how?

Q202 So what is the difference between thinking about, or cyber security practices at home compared to at work?

Q203 So how much does your workload and the pace of your workplace vary over time?

Q204 So you work for a bank – does that bring with it any special responsibilities and roles?

Q205 Do you know who the cyber security team are the bank? Or how to find them or contact them?

Q206 Do you think the Cyber-security team are good at what they do?

Q207 Do you get a lot of emails that aren’t really relevant to you? Or are trivial?

DigiTally: Piloting Offline Payments for Phones

Khaled Baqer
Computer Laboratory,
University of Cambridge
Cambridge, UK
Khaled.Baqer@cl.cam.ac.uk

Ross Anderson
Computer Laboratory,
University of Cambridge
Cambridge, UK
Ross.Anderson@cl.cam.ac.uk

Jeunese Adrienne Payne
Independent Consultant
Cambridge, UK
jp@jeunesepayne.com

Lorna Mutegi
Strathmore University
Nairobi, Kenya
lmutegi@strathmore.edu

Joseph Sevilla
Strathmore University
Nairobi, Kenya
joe@strathmore.edu

ABSTRACT

Mobile payments support a range of services in many less developed countries including everyday payments, migrant remittances, credit, tax collection, and welfare benefits. These services depend entirely on the mobile phone network as their carrier, so they stop where the network does. This leaves millions of the very poorest people stranded – people living in remote areas where there is little to no network service. It also leaves urban users at the mercy of network congestion.

We developed a prototype system, DigiTally, which lets users make offline payments by copying short strings of digits from one mobile handset to another. Offline payments are already used for electricity (both in prepayment meters and pay-as-you-go solar); can we extend them into a general-purpose payment system, to increase service resilience in the face of network congestion or outage, and provide service to currently excluded areas?

We report the results of a preliminary study with an early prototype of DigiTally, tested on participants from a university in Nairobi (Kenya). The code-sharing process presented a possible usability challenge. To explore this and other aspects of an early prototype, DigiTally was introduced to Kenyan participants in order to resolve any major issues before a later field trial.

We discuss the lessons learned from our field visits and initial evaluation; we hope that this contribution is helpful for researchers and policy makers interested in mobile payments and financial inclusion. We also present our findings and observations. We found that, although offline payments involve copying codes in both directions between the payer's phone and the payee's, the extra workload was acceptable to most users.

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

1. INTRODUCTION

Mobile payments have transformed the lives of millions in less developed countries, bringing a means of exchange and a store of value to people who did not previously use a bank account or who lived far from any bank branch. They run on top of a mobile phone network; a user can typically load their phone with credit at the same agent where they buy airtime, and can send money to other service users. Over 200 such services have been launched worldwide and about 20 have achieved serious scale[1]; one pioneer was Kenya's M-Pesa [11], operated by the local phone company Safaricom. The initial killer application was migrant remittances, but M-Pesa is now very widely used for everyday purchases as well as specialist applications from paying pensions and government farm subsidies to collecting business taxes [12].

Because of the strong positive effect on development, the Bill & Melinda Gates Foundation called for innovations that could increase the uptake of mobile payments[2]. One of the largest impediments is that current systems operate entirely online; both the payer and the payee have to be able to communicate with the payment system server for a payment to be completed. This excludes millions of people living in remote areas with no network service; such people make up 10-40% of the population depending on the country. It also makes payments harder in the event of network congestion (we have observed 30-second delays in down-town Nairobi). Additionally, where the payment service operator is not the same firm as the mobile network operator, charges become an issue.

The main contribution of this paper is to describe a preliminary study that took place at Strathmore University in Nairobi, in September 2016. We set out to establish whether DigiTally was usable in three different environments: a coffee shop, a campus bookshop, and a cafeteria, and by students from a range of backgrounds. These students were experienced users of M-Pesa and thus able to compare it

[1]As shown on GSMA's 'Mobile for Development' website (using the *Mobile Money* filter): `http://www.gsma.com/mobilefordevelopment/tracker`
[2]Enable Universal Acceptance of Mobile Money Payments: `https://gcgh.grandchallenges.org/challenge/enable-universal-acceptance-mobile-money-payments-round-14`

with DigiTally; we also got assessments from the checkout staff. Further contributions are as follows: the paper analyses a security technology and offers lessons learned from a preliminary usability study. The paper should be of interest to researchers interested in development, and also to those interested in evaluating payment systems; both communities can benefit from our insights. Documenting the study in one publication should help them and others interested in this type of research.

We discuss the background and related work in section 2. We describe the technology in section 3, and we discuss our method and results in sections 4 and 5. We describe our observations in section 6, and provide more discussion in section 7. Finally, we present our conclusions in section 8.

2. BACKGROUND AND RELATED WORK

Modern mobile payment systems in less developed countries rely on encrypted Short Message Service (SMS) messages or Unstructured Supplementary Service Data (USSD) sessions, supported by a Subscriber Identity Module (SIM) card issued by a phone company. Examples include M-Pesa in Kenya [11] and bKash in Bangladesh [21, 5]. The SIM can verify a user PIN and contains keys for authenticating the customer; it can thus establish secure communication with a payment server. The server keeps everyone's transaction histories, just like in a conventional bank. Customers can check their balance and make payments using menu options on the phone's screen, displayed by the trusted SIM card. The phone number doubles as a bank account number. Cash-in and cash-out services are provided to customers by the mobile payment operator to facilitate exchanging fiat currency into electronic float and vice versa. These services are operated by a network of mobile money agents, who earn commissions based on the services provided (e.g. withdrawing funds from the system incurs a fee). To make a payment, a customer first enters the merchant's phone number (shops have numbers prominently displayed, with large shops having short codes). The customer must then enter the amount, followed by their PIN to authorise the transaction. A payment message is sent to the server; if the funds are available, they are transferred, and the merchant is informed. The merchant knows when the money has been received because their phone pings.

Jack and Suri discussed the economics of M-Pesa in [12], highlighting issues that affected system uptake, including liquidity and network reliability. Zimmerman and Baur discuss the challenges facing financial inclusion efforts [23], including network coverage and reliability, liquidity, complexity of user interfaces and payment processes, the lack of dispute resolution, and the lack of customer protection against fraud. Dupas *et al.* report that a significant proportion of the participants in their study listed fraud (embezzlements), unreliable services, and transaction fees as issues [9].

The goal of our work is to tackle the network coverage and unreliable services problems by processing payments *offline* reliably and deterministically, and to simplify the user interface as much as possible by mimicking familiar mobile payment systems. Furthermore, we aim to decrease transaction fees to encourage users to process transactions electronically. The purpose of this paper is to investigate possible usability challenges to be addressed in subsequent versions of our prototype. Offline payment systems have already been implemented, such as Geldkarte[3] in Germany and Net1/UEPS [1] in South Africa, but require dedicated devices or unfamiliar hardware that can be costly to replace if lost or stolen. Similar systems might perhaps be implemented on modern smartphones if both the payer and the payee had them. However, most users in less developed countries still use feature phones that do not have NFC, Bluetooth, or cameras; these phones cannot communicate data automatically in the absence of a network. We also minimise the assumptions we make regarding what features are available on users' phones.

Our goal is to design a system that operates without relying on such features, in case some of them are inoperable. For example, if we rely on a camera and it is damaged, the entire system is useless until the user has their camera fixed. By operating within these constraints, we are able to design a system that works on all low-end mobile devices; our project aims to provide a solution for the poorest demographics.

The use of feature phones emerged as a requirement during trips to rural Kenya. Users in Busia county, for example, specifically requested solutions that work on feature phones. One person commented: "*Don't give us [systems that work on] smartphones. We don't have those things and we don't know how to use them.*" When asked what kind of phones users owned, all but one person (in a group of more than 20), held up a feature phone; the non-feature phone was an old BlackBerry. Reliance on feature phones (called '*kabambes*' in Kenya) is due to both battery life and cost (in terms of price as well as maintenance). With such constraints, the only way to transfer information offline is for one phone to display it, and for a human to type it into the other phone (a similar approach used in device-pairing methods). Kainda *et al.* [13] looked at the tradeoff of usability and security with regards to different device-pairing methods, using out-of-band channels, as it applies to various device usage or capability restrictions. Their results show that typing strings (*"copy & enter"*) ranked first in terms of the aforementioned tradeoff. They also recommend that system designers take into account factors that affect that tradeoff including user conditioning, user motivation, security failures, and attentiveness. Moreover, transferring value by copying digits is well established in other applications, such as prepayment electricity meters [2].

3. TECHNOLOGY

An early design goal we set was that DigiTally must not require users to operate any unfamiliar hardware, and that the transaction flow should be as close as possible to the familiar one. This leads immediately to the challenge of programming feature phones: different operating systems make it difficult to implement and test applications for various models, leading to large costs for maintenance and support.

To minimise these costs within our constraints, the practical approach is to program the SIM card in the user's phone. SIMs were designed to host multiple applets in secure containers, which can prevent one applet from accessing other applets' data. A SIM provides a secure environment that we can control and that is compatible with all mobile devices adhering to the Global System for Mobile Communications (GSM) and smartcard interoperability standards. This

[3]Geldkarte website: `https://geldkarte.de/`

Figure 1: An overlay SIM (top) and a regular SIM (bottom). The top part of the overlay SIM can be peeled off and stuck on top of a regular SIM

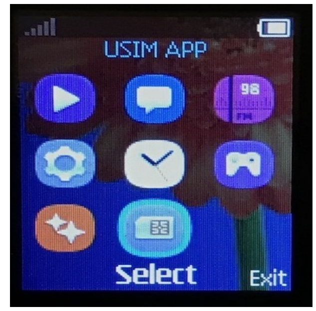


Figure 2: Selecting DigiTally applet from the feature phone's application menu

Figure 3: DigiTally user interface on a feature phone

means that we can program our system on a SIM, and insert it into any device that accepts a SIM. The user can move the SIM from one device to another, allowing for portability. SIMs also provide valuable built-in features, including atomic operations and rollback mechanisms, as well as the ability to store secure tokens and cryptographic keys in a tamper-resistant chip.

Feature phones normally have a single SIM slot already taken by the SIM issued by the Mobile Network Operator (MNO). MNOs do not generally let anyone else program their SIMs, but there is a new technology that can be used to circumvent this restriction. This is the overlay SIM (or sticker SIM): a SIM card only 120 microns thick that can be inserted between the regular SIM card and the phone. Figure 1 shows an overlay SIM and a regular SIM. Overlay SIMs were developed to support low-cost mobile roaming. They can also be used as a proof-of-concept prototyping environment, and to bypass MNOs' restrictions on devices if necessary. We used the programmable overlay SIM as a regular SIM, inserted in the single SIM slot. The same overlay SIM can also be stuck on top of an existing non-overlay SIM, and our system will work seamlessly on the device while allowing it to still work as a phone[4].

3.1 Overview of DigiTally

We designed and developed DigiTally as a Java card applet accessed through a user's phone (Figure 2). This applet can be loaded on an overlay SIM or a regular SIM. We chose to use a regular SIM for this study (for reasons we discuss in section 4.4 item 2). As discussed earlier, having an overlay-SIM ready applet enables deployment on feature phones even if the MNO chooses not to install DigiTally on their own regular SIM.

Rather than asking participants to use their own phones, we provided participants with dedicated phones (Nokia 130), preloaded with a SIM that included the DigiTally applet. The applet includes a menu that mimics existing mobile payment systems to capitalise on users' familiarity with them. The applet offers the user options such as `Balance`, `Send Money`, and `Receive Money` (Figure 3). Figures 8 and 9 illustrate the steps required to complete a transaction.

[4]Overlays SIMs are used in China for roaming purposes, and in Kenya by a local Bank to provide financial services to their users and to break local phone-payment monopolies. Overlay SIMs work on standard carriers' SIMs.

The icon for the applet displayed in Figure 2 is enforced through the mobile OS. When a single SIM is present, selecting the SIM icon will display the applets that reside on that SIM. When multiple SIMs are present, such as in the case of using an overlay SIM on top of a regular SIM, selecting the SIM icon directs the user to two different menu options that represent the applet names in the two SIMs. Note that there could be more than one menu option displayed after selecting the SIM icon, if one or both of the SIMs offer more than one applet.

There are two main differences between traditional mobile payment systems and DigiTally. The first difference is that the SIM in an offline payment system stores the *balance* as a value counter, and the payment protocol changes this local balance in the SIM, whereas a traditional mobile payment system merely sends requests to a backend system to process operations on the user's balance. The second difference is the `Receive Money` option: in the absence of an online payment server, the recipient must be involved in completing a transaction. In traditional mobile payment systems, the recipient is not actively involved; they merely get an SMS saying how much money has arrived and from where. In offline payments, both phones must be involved in a transaction. The protocol is designed so that the parties learn immediately whether the transaction has failed or completed. Note that DigiTally does not require data to be sent over the network and can, therefore, work completely *offline*.

The technical description of the DigiTally protocol [4] describes the cryptography and other security mechanisms ([4] does not include a description of the design or implementation of DigiTally). For the purposes of this paper, we focus on the codes exchanged by the users to complete transactions. These are Message Authentication Codes (MACs) that establish agreement between the payer and the payee on the transaction details, namely the payer, the payee, and the amount. These codes are computed using secret keys kept in their SIM cards based on the transaction inputs and previous history, and are designed to prevent transaction manipulation, replay, or guessing.

3.2 DigiTally Codes

We now outline the stages required to complete a DigiTally transaction. There are two codes involved in completing a transaction. To simplify the discussion, we will assume that Alice is paying Bob.

1. $Code_1$: After the payee B (Bob) has entered into his SIM the transaction amount X and the phone number of the payer A (Alice)[5], his SIM generates a random nonce (N_B), and then computes a MAC on A, B, N_B, X and the log of previous transactions ℓ between the two parties[6]. This MAC and the nonce together make up the 8-digit $Code_1$, which is shown on the payee's device (Figure 4). Alice similarly enters into her phone the amount X and Bob's number B[7]; it prompts her for $Code_1$, which she enters (Figure 5). If the two parties agree on X, ℓ and each others' identity, then Alice's phone accepts $Code_1$. If there's a disagreement – whether due to attempted cheating or an honest mistake – her phone will generate an error[8].

2. $Code_2$: If $Code_1$ is correct, Alice's SIM card decrements her account's available balance by the transaction amount X and generates $Code_2$ to authenticate the transaction. $Code_2$ is also 8 digits long; it consists of a 4-digit nonce N_A generated by Alice's SIM and 4 digits from a MAC on A, B, X, ℓ and N_B. Alice then shows or tells $Code_2$ to Bob (Figure 6). He enters $Code_2$ into his phone, and, if it is valid, his SIM increments his balance by X and a transaction log is displayed on his phone (Figure 7). A similar transaction log is shown to Alice to confirm the completion of the transaction (decrementing her balance by X).

This is the simplest payment protocol we could devise that enables value to be transferred from Alice's card to Bob's by copying 8 digits in one direction and 8 digits in the other. Its security is analysed and discussed in [4]. Here, our focus is usability; the STS prepayment electricity meters widely used in Kenya transfer value by means of a 20-digit number, presented as five groups of four digits [2]. A householder buys codes from an ATM or sales agent and copies them into their electricity meter; codes can also be bought online, using

[5]In our trial, user identities are randomly generated numbers, each simulating a phone number.
[6]If no previous transactions exist, then the first transaction initialises the relationship between the two parties.
[7]Alice and Bob can pick each others' names from a menu. The first transaction stores the contact's details, and subsequent transaction can later retrieve information from the locally saved contacts (on the phone or SIM).
[8]For example, if Alice enters \$4 on her device, and Bob enters \$5, then Bob's $Code_1$ will generate an error on Alice's device. The same thing happens if the wrong phone number is selected.

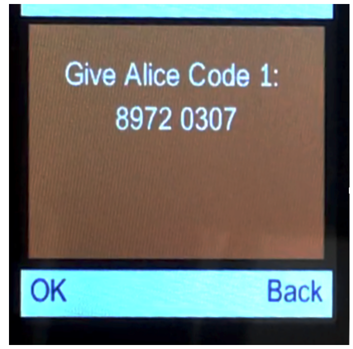


Figure 4: Bob's phone displays $Code_1$ that must be given to Alice

Figure 5: Alice enters $Code_1$ into her phone

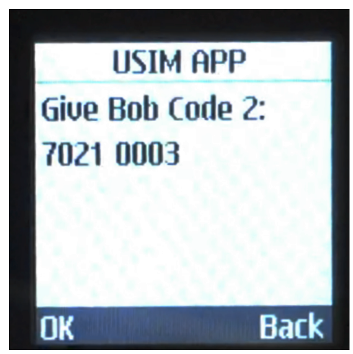


Figure 6: Alice's phone displays the response code ($Code_2$), given to Bob to authorise the payment

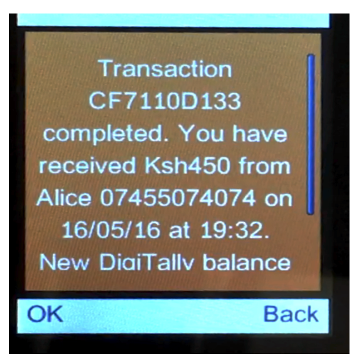


Figure 7: Bob's phone displays the transaction log after accepting $Code_2$

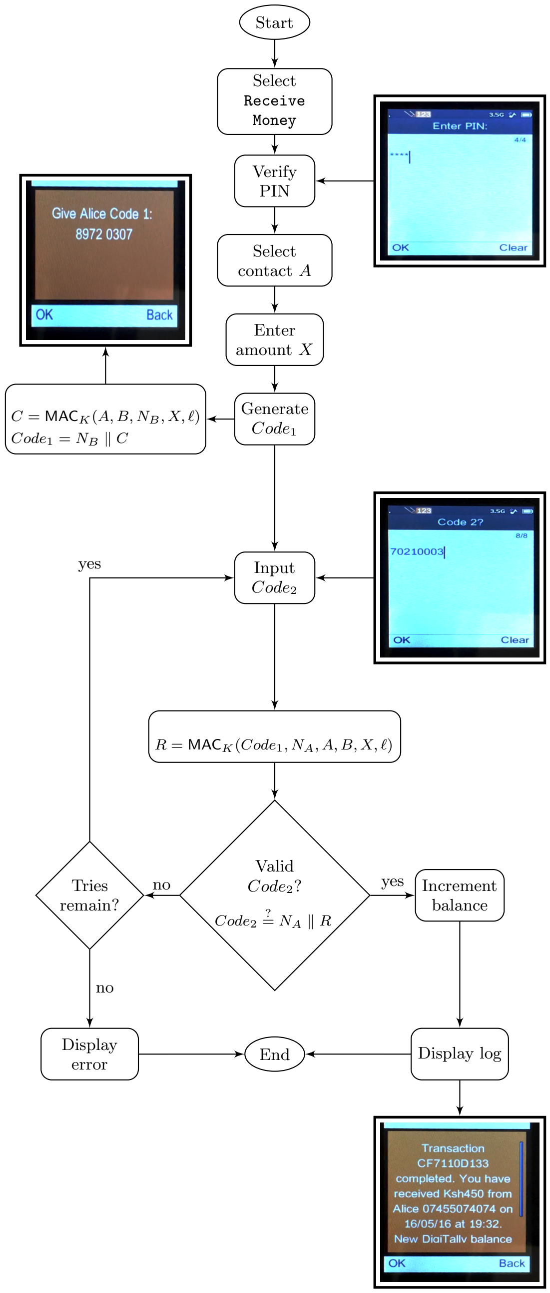


Figure 8: Recipient's (Bob B) transaction steps

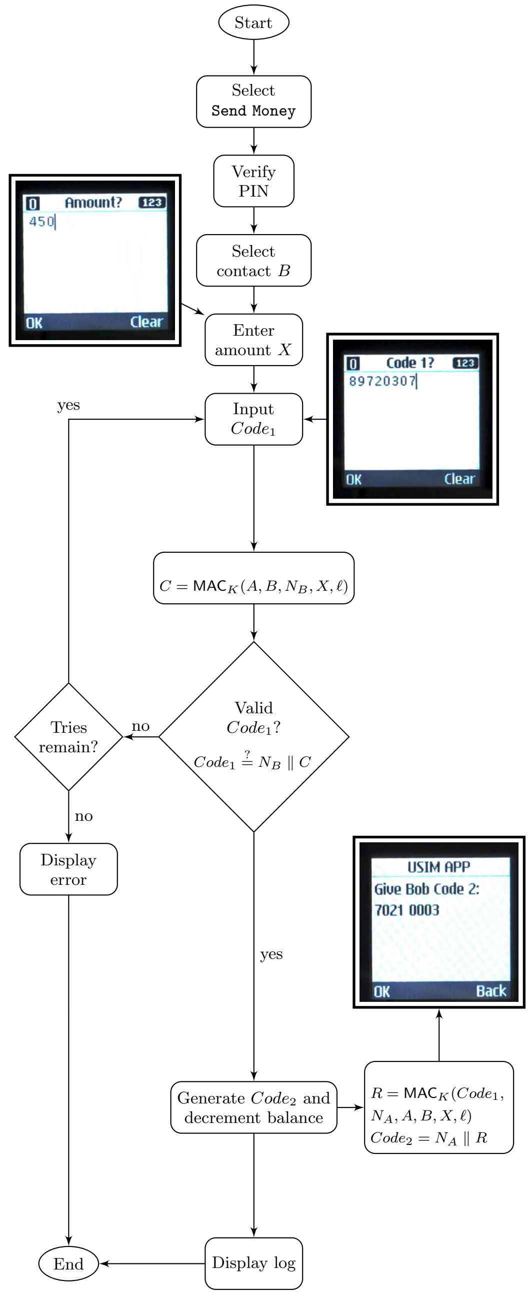


Figure 9: Payer's (Alice A) transaction steps

M-Pesa. Thus, we know that our prospective customers can copy digits, and that even illiterate people can use a phone. The object of the study is to establish whether the DigiTally transaction flow is usable.

Some care is necessary to ensure robustness in the face of mistakes. The DigiTally implementation includes various segments that execute atomic operations (enforced by the Java card platform) to ensure that critical components of the payment protocol are fully executed (or reverted to their initial state if a problem occurs). This provides the ability to reliably create *checkpoints*. We checkpoint the transaction when $Code_1$ is successfully generated so that a transaction can be recovered if, for example, a merchant accidentally generates a second, new $Code_1$ before the customer replies to the first one. Additionally, as Alice's balance is decremented prior to displaying $Code_2$ on the payer's device, we have to ensure that if she is interrupted (e.g. by a flat battery), her SIM can still retrieve $Code_2$ later ($Code_2$ is saved in the final checkpoint of the transaction, and can be retrieved from the "last transaction" log). Checkpointing was also tested (see section 6.2).

DigiTally codes do not have to be kept secret, since the codes can be used only once and only by the payer and payee in a specific transaction. There is no added risk if other users observe the transaction or overhear the codes being exchanged verbally. This is in contrast to mobile payment systems that rely on secret codes: codes that compromise the users' security if intercepted, and increase the burden on the payer to deliver them to the recipient out-of-band. Users of feature phones can't use a secure messaging service, so probably have to use voice or SMS and hope for the best. As well as the (small) technical fraud risk, such codes can create anxiety and make dispute resolution problematic.

4. METHOD

The usability evaluation of DigiTally was conducted by the primary researcher from the University of Cambridge (UK) and researchers from Strathmore University (Kenya), in early September 2016. Participants from Strathmore University (hereafter referred to as 'the university') performed real-world transactions using DigiTally and then answered questions about their experience. As well as open ended-questions, we make use of the System Usability Scale (SUS), which is a standardised tool widely used for measuring usability. We give details below.

4.1 Participants

We wanted to test the technology as part of a pilot, and so we needed participants who could reveal weaknesses in our design before testing it as part of a field study with a representative sample. The aim was to maximise usability before introducing DigiTally in a larger-scale field test with target users. As an initial trial to establish any major issues with the DigiTally system itself, we recruited Kenyan participants from the university.

The participants were recruited following their registered interest in the study in response to an advertisement. Potential participants were interviewed with an aim to achieve diversity in terms of demographics, and to establish that participants would be able to give us detailed feedback to inform major re-designs before a field trial with a representative sample of rural users.

Twelve students and seven merchant staff were recruited as participants to use DigiTally for five days. Merchants were recruited from the university's cafeteria, bookshop, and a local coffee shop. According to Sauro [18], Vizri [22] and Nielsen and Landauer [16], very few participants are needed for early phase usability studies such as this, since adding more users tends to uncover the same issues, with 'five' often referenced as the "magic number" of participants. Confidence intervals may be wide as a result, but the average SUS score should be stable [17].

For wider coverage, we included a greater variety of participants, which required larger numbers than the proposed "magic number 5". This is in line with more cautious estimates of 10-20 participants for a usability test (e.g. Faulkner [10] and Macefield [15]).

In total, we had 19 participants for this preliminary study. There were 7 female and 5 male students, who were studying a range of topics in different faculties, including finance, law, and information technology. As for the merchants, the cafeteria included 1 female and 2 male staff members; the bookshop had 1 female and 2 males; and the local coffee shop had 2 females (only one of whom was responsible for processing DigiTally transactions). Each merchant was given one feature phone to process DigiTally payments; staff members in the cafeteria and bookshop shared the device to process payments, while the coffee shop had one person operating the feature phone.

Because we did not provide cash-in and cash-out services (to convert physical currency to DigiTally balances, and vice versa), participants were motivated to use DigiTally to spend the balance they had in their SIMs.

4.2 Evaluation Materials

Questionnaires and surveys are one of the most widely used methods for measuring attitudes [8]. They often involve asking participants to select one of a number of points on a rating (Likert) scale [20]. Good scales are valid (in that they represent the intended construct), and reliable (giving measurement consistency) [7, 20]. Standardised usability questionnaires and surveys are preferred because they are quantifiable, are economical to reuse, and allow for replicability of findings [20]. The most popular standardised survey for measuring attitudes towards usability is the SUS, favoured due to its brevity and being free to use [19]. The items in this survey factor into two sub-constructs of usability: system learnability (ease-of-use for new users), and system usability (defined as ease-of-use more generally) [19].

It was important to use feature phones for this study rather than the smartphones that many students normally use. Participants were thus given identical feature phones to test the basic usability of the technology first; a field test with representative users in rural Kenya who already use feature phones would be the next step.

We asked participants to complete a short pen and paper survey, which included the SUS, to make an initial assessment of DigiTally. Here, we asked participants for three free-text responses, to get insights into how they viewed DigiTally and where we might need to improve. These were:

1. What did you like about DigiTally?
2. What did you dislike about DigiTally?
3. Additional comments

4.3 Procedure

Students and on-campus merchants at the university were invited to try DigiTally and to give their opinions of their experience of the system. This involved using dedicated feature phones to transact using the DigiTally applet, which was installed on each phone's SIM. The students' phones were preloaded with a DigiTally balance of Ksh 2000 (about $19.50) that could be used to make transactions with merchant participants who accepted DigiTally. The merchants (cafeteria, bookshop, and a local coffee shop) were given phones that had a zero balance on their DigiTally applet.

Before the trial, participants gave informed written consent and were shown how DigiTally works. After the trial, participants were asked if they were willing to complete a survey about their experience, which was entirely voluntary. This survey consisted of the SUS and the three open-ended questions outlined in section 4.2 (Evaluation Materials).

Each day towards the end of business hours, the researchers visited each merchant to check their balances and reimburse them in cash the amount they had in their DigiTally balance (after transferring the DigiTally balance to another phone, and thus resetting a merchant's balance to zero at the time of reimbursement). Each merchant signed a form indicating the date and amount of reimbursement. The reimbursement forms were necessary to create a detailed accounting of the money for our funders and auditors.

4.4 Ethics, Data Collection and Privacy

This study was approved by ethics committees at both universities involved in this research (one in the UK and the other in Kenya). To address the various concerns raised by the ethical review boards in both universities, we changed the design of the study to minimise any possible risks:

1. **Financial risks**: to address the risk of possible loss of personal funds, we provided the participants with a preloaded amount into their DigiTally applet (as explained in section 4.3). We decided not to provide cash-in and cash-out services to eliminate the possibility that participants would use their own funds to increase their DigiTally balances.

2. **Privacy risks**: we were asked for assurance that the overlay SIM would not compromise the privacy of participants by interfering with the participants' regular SIMs. Although we initially intended to use an overlay SIM on top of a regular SIM, we decided instead to use the overlay as a regular SIM and provide users with dedicated devices. The reason for this decision is that *proving* that an overlay SIM is not compromising users' privacy was out of the scope of this project. Note that the overlay SIM (and the DigiTally applet) does *not* compromise users' privacy in our implementation, since it operates as an independent sandboxed applet and does not interfere with other applets' data if there are any installed.

Therefore, we consider the preliminary study to have no significant ethical implications. There were no known risks to participants (financial, psychological, emotional, or physical). Participants were required to provide informed consent before starting the study. The informed consent form outlined what the study involved from the participant, how long it would take, our commitment to confidentiality, and their rights as a participant. Participants were reminded, verbally and in the consent form, that their participation was entirely voluntary, and that they may choose to not answer particular or all questions and may withdraw from the study at any point without having to provide a reason, and without fear of penalty from the researchers.

This study did not involve deliberately deceiving participants in any way. This study also did not involve recruiting participants from vulnerable groups, such as children, patients, people with learning disabilities, people engaged in illegal activities, or people in custody. Participants were fully debriefed and their questions were answered throughout the study. Additionally, participants were given contact information for all of the researchers from both universities should they have any further enquiries.

The data we used for our evaluation includes the survey results, as well as the performance and transaction data processed by the SIM. By collecting these data, we were able to analyse performance and usability issues as we discuss in Results (section 5). We collected the following:

1. Error rates (code-entry errors, and wrong PIN input)

2. Number of transactions

3. Number of attempts to unlock the SIM

4. Total amounts for all transactions (spent and received)

5. Transaction duration times

The SIM also stored the full log of the last transaction, which is overwritten when a new transaction is successfully processed[9]. This log contains $Code_2$ which can be retrieved after the transaction is completed. Recall that $Code_2$ can be used only once and with the right recipient (the identity of the recipient is one of the inputs required to generate the code), but a user might want to retrieve $Code_2$ in case their battery goes flat before exchanging the code with the recipient.

The collected data were stored in the SIM card (a trusted tamper-proof element). DigiTally does not leak any information outside this trusted element. Therefore, performance data are inaccessible without the correct authenticator. Only pre-configured devices, programmed to provide the correct authenticator, had access to the SIMs to obtain the performance measurements. Each SIM was programmed to lock itself after a certain number of failed attempts to access the contents of the trusted tamper-proof environment. At the completion of the preliminary study, we retrieved the phones from all the participants to extract the performance data from the SIMs and save it for later analysis. All the stored data were encrypted after extraction.

With the exception of participant names, no Personally Identifiable Information (PII) were collected in the study (such as date of birth, national ID numbers, etc.), as they were irrelevant to the scope of the study. We carefully explained to all participants what data were being collected, by demonstrating a full transaction on two phones and explaining the performance variables to be measured for the purposes of the preliminary study.

5. RESULTS

The following sections report on the actual error rates and transaction times, as well as the SUS results. We present descriptive statistics to help summarise the data and to show any emerging patterns. We make no claims about statistical significance, which would be reserved for inferential statistics. This would not be appropriate since the tested population do not represent the intended users of DigiTally, and were recruited in order to reveal any major usability issues with the technology itself. We thus make no claims about generalisability of the results.

We also discuss the open-ended results and the lessons learned through the feedback obtained from the users. We discuss

[9]This log is informational only, and is different from the cryptographic parameters securely stored on the SIM and used in the payment protocol as discussed in section 3.2. Refer to [4] for more details.

the features that participants considered to be the most important, and indicate issues that may need to be re-evaluated to improve usability.

Note that there is a difference between the total number of transactions completed by students, and the total number of transactions completed by merchants. This is due to the fact that students moved funds between their phones, perhaps to test the system and in some cases to demonstrate it to their friends; moving funds between devices back and forth did not incur any costs in terms of transaction fees.

5.1 Errors and Speed

Students completed an average of 23 transactions. The highest number of transactions completed by one participant was 30; the lowest was 8. For merchants, the average number of transactions was 59. The highest number of transactions completed by one merchant was 61; the lowest was 58. Tables 1 and 2 display a summary of the ***Error Rates*** and ***Time on Task***, which should be considered against the number of transactions each participant successfully completed.

5.1.1 Error Rates

DigiTally captured the number of errors participants made while trying to complete the task. Student participants made the most errors when entering the first code ($Code_1$) presented to them by the merchant. The number of errors here ranged from 0 to 9, with 7 of the 12 student participants making this error. These were non-critical errors since they do not prevent successful completion of the transaction. However, future trials will need to test the ease with which users can recover from this and other errors. For $Code_2$, the number of errors ranged from 0 to 8, with 5 of the 12 student participants making this error.

The merchant participants made the most errors when entering $Code_2$ which is required to authorise and complete the transaction. The number of errors here ranged from 4 to 6. Merchant $M2$'s $Code_2$ errors were higher because they included errors experienced during training. We adjusted the number in the table to reflect the errors experienced during the study without those experienced during the training. Merchants made no $Code_1$ errors; this type of error would occur if they used DigiTally to make payments instead of receiving (the merchants made a payment using DigiTally for the daily reimbursement transactions, see section 4.3).

We include PIN errors in Tables 1 and 2 for completeness. However, we do not consider entering PINs to be a great problem, at least no more than in any other system that uses them. For clarity, we report (in brackets) the proportion of errors relative to the number of transactions completed as a percentage (for non-zero values).

5.1.2 Time on Task

DigiTally recorded the time on task for each participant. The transaction timer starts when a menu option is selected (`Send Money` or `Receive Money`), and the timer is stopped before the transaction log is displayed to the user (the transaction log includes the transaction's duration time).

The average time for students to complete a transaction was 36.4 seconds. Their average completion times ranged from 24.2 seconds to 54.3 seconds. It is worth noting that a large number of errors by a participant did not necessarily translate into longer average time spent on transactions. For the merchant participants, the average time to complete a transaction was 51.23 seconds. Their average completion times ranged from 40 to 69.8 seconds. Merchants' transaction times are larger since a merchant can start a transaction, give $Code_1$ to the payer, then complete a few tasks to serve the customer until $Code_2$ is entered. Such tasks include having to prepare food at the same time as processing DigiTally payments, which is the case for merchant $M3$. For merchants dedicated to processing payments (cashier roles), the transaction times are lower ($M1$ and $M2$).

5.2 SUS Results

The SUS is not diagnostic: it will not reveal specific problems, but it does give an idea about overall *ease-of-use*, and whether significant changes might be needed.

To give a clear idea of the results, the SUS is calculated so as to provide a score out of 100. However, the SUS Score is not a percentage. A score of 68 actually falls at the 50th percentile (i.e. the average SUS score is 68). If the score is below 68, there are likely to be serious usability problems that need tackling. A score of 80.3 or higher is ideal.

The average SUS score for DigiTally was 78.8, which is considered 'Good', and would be given a 'B+' grade. The lowest score was 50, which is considered 'Poor', and is equivalent to a 'F' grade. This score was given by a merchant and was the only score considered 'Poor' by SUS standards. The highest score was 100, which is the best possible, and is equivalent to an 'A+' grade. Eight participants gave the equivalent of an 'A+' grade.

For merchant participants, the average SUS score was 71.4, which is considered 'Good' and equivalent to a 'C+' grade; for student participants, the average SUS score was 83.1, which falls just short of being considered 'Excellent', and is equivalent to an 'A' grade.

Although the SUS is intended to be a measure of ease-of-use, Lewis and Sauro argue that it can also be used as a measure of *learnability* (using items 4 and 10 of the SUS) [14]. As with calculating the SUS score, learnability can be calculated to give a score ranging from 0-100. The average learnability for this initial trial was 82.9, which falls just short of 'Excellent' and would be given an 'A' grade. The remaining 8 items are what Lewis and Sauro call a measure of *usability* [14]. For the current initial trial, this score was 77.8, which is also considered 'Good' and would be given a 'B+' grade.

5.3 Responses to Open-Ended Questions

Participants' answers to the three open-ended questions were categorised into several themes. This qualitative data analysis involved two researchers independently coding user comments to identify common themes, supported by quotations. These researchers then came together to assess agreement and categorised themes based on a well-established definition of usability (ISO-9241), which consists of effectiveness (usefulness), efficiency (ease-of-use), and satisfaction.

5.3.1 Perceived usefulness

1. **Money saving**. Most participants mentioned the benefit of there being no transaction fees. They pointed

	No. of transactions	PIN errors	$Code_1$ errors	$Code_2$ errors	Total code errors	Average time (seconds)
S1	30	2	0	8 (26.7%)	8 (26.7%)	30.9
S2	28	3	9 (32.1%)	0	9 (32.1%)	24.4
S3	18	3	2 (11.1%)	8 (44.4%)	10 (55.6%)	28.1
S4	22	0	9 (40.9%)	1 (4.6%)	10 (45.5%)	44.9
S5	29	1	1 (3.5%)	0	1 (3.5%)	24.2
S6	26	1	0	1 (3.9%)	1 (3.9%)	54.3
S7	26	0	1 (3.9%)	0	1 (3.9%)	50.9
S8	28	0	5 (17.9%)	0	5 (17.9%)	32.4
S9	10	0	0	4 (40.0%)	4 (40.0%)	28.8
S10	29	1	0	0	0	37.1
S11	8	2	0	0	0	42.1
S12	22	0	5 (22.7%)	0	5 (22.7%)	38.9

Table 1: The frequency and types of errors made by student participants and the percentage of each participant's transactions that their errors affected

	No. of transactions	PIN errors	$Code_1$ errors	$Code_2$ errors	Total code errors	Average time (seconds)
M1	58	2	0	4 (6.8%)	4 (6.8%)	40
M2	61	0	0	6 (9.8%)	6 (9.8%)	43.9
M3	58	2	0	4 (6.8%)	4 (6.8%)	69.8

Table 2: The frequency and types of errors made by merchant participants and the percentage of each participant's transactions that their errors affected

out that this would be useful for those in poor communities, as well as being attractive to price-sensitive customers and merchants. Aside from helping users avoid transaction fees, one participant mentioned that it also meant they did not have to use a smartphone, making DigiTally even more cost-friendly.

2. **Network independent transactions (interoperability)**. Many mentioned the benefit of not having to rely on network coverage. For some, this made the DigiTally transaction process seem more reliable. It also makes it more predictable as users do not have to wait for a confirmation SMS, which with M-Pesa can take up to 30 seconds.
3. **Security**. The general consensus was that DigiTally seemed very secure. The codes were a major factor behind this perception. While recognising this benefit, a few participants suggested that the codes were too long, and recommended that the developers consider shortening the codes.
4. **Money tracking**. Participants liked being able to review their last transaction and balance. For some, however, this was not enough. It was suggested that this feature would be more useful if the user could review all previous transactions.

5.3.2 Perceived ease-of-use

1. **Ease**. Participants perceived DigiTally to be simple and easy to use. Some clarified that DigiTally was easy to use *after* a learning period. One merchant described the process as cumbersome; this merchant was both serving customers and taking payments. Other merchant participants were either less busy or were dedicated cashiers. One participant also stated that DigiTally might be harder to use for the elderly and those with poor eyesight.
2. **Learnability**. DigiTally was most often praised for its learnability; in general, participants felt like DigiTally was easy for a first-time user to understand. Even the busy merchant ($M3$) who had complained that the system was cumbersome said she had no difficulty training a staff member to use it. The same merchant also warned that learnability might be lower for some, including the target population, where there might be more illiteracy and less education more generally. Another merchant was curious to know how money would be deposited in a production DigiTally system (this merchant is already an M-Pesa agent).
3. **Speed**. Participants found transacting with DigiTally to be relatively fast, once they knew the process. Two out of the eight merchants said that DigiTally had too many steps and was time-consuming compared to cash or M-Pesa. Requiring as much effort from the merchant as from the customer to give and receive codes was problematic when the merchant had a lot of customers and needed to perform other tasks at the same time. One merchant also did not like having to search for the customer in their contacts list. One participant pointed out that the speed of transaction would be less of an issue in rural areas.
4. **Errors and recoverability**. One participant stated that it is hard to make errors that would cause the user to lose money, and that they found errors easy to rectify. Many others did not agree. In general, error reset was considered too difficult. Although the process for error recovery was perceived as cumbersome, none of the users had to go through it during the trial. See section 6.2 for more details.
5. **Cashlessness**. Two participants noted that by using DigiTally they would not need to carry cash around. This was considered convenient. It was also noted that

merchants do not have to find the exact change when dealing with customers using DigiTally. This benefit is shared with other electronic payment systems; Kenyans already like M-Pesa as it dispenses with the risk and the inconvenience of cash.

6. **Codes**. Although many did not have much of a problem with the codes, finding them short enough and necessary for security, some stated that they did not like them, mostly because the code or the process was too long or awkward. Having to exchange and input codes was also perceived as an opportunity to make errors. On the other hand, one of DigiTally's perceived advantages was the deterministic nature of the transactions, because exchanging the codes provided immediate feedback that a transaction was completed, without needing to wait for an SMS.

7. **Distance**. Sending money to a remote payee was identified as a potential problem. It was pointed out that errors could be more likely and recovering from them would be harder if users were trying to complete a transaction at a distance.

5.3.3 Satisfaction

1. **Likeability**. Participants liked using DigiTally, using words like 'happy', 'good', 'best', 'seamless', 'enjoy', 'enthusiasm', 'encouraging', 'smart', and 'satisfied' to describe their experience. Comments suggest that likability could be improved by targeting issues associated with the length of the codes and with recovering from errors.

2. **Other services**. In addition to a merchant who wanted to know how to do cash-in and cash-out transactions, one student participant also indicated that they would like to be able to do this using DigiTally. We did not provide cash-in and cash-out services during the trial.

6. OBSERVATIONS

We describe in the following sections our observations during the preliminary study. These observations were documented during regular visits to merchants to answer their questions, and during daily visits to reimburse merchants for their DigiTally transactions (as discussed in section 4.3). We highlight two important observations, visual cues and the error-recovery process, and discuss the lessons learned.

6.1 Visual Cues

We observed that participants chose to display the codes so that the other participant could see them and enter them into their own device. None of the participants indicated that they exchanged the codes verbally, and some stressed that visual exchange is easier. In crowded areas, a verbal exchange of the codes could lead to misheard digits. The cafeteria is a crowded environment with long lines of customers, where a window separates the cashier and the customer and a small opening at the bottom of the window allows exchanging cash or passing cards (Figure 10). In this environment, we observed multiple DigiTally transactions where the only verbal exchange was to acknowledge that a transaction was completed by saying "*it's OK*", and in some cases just a

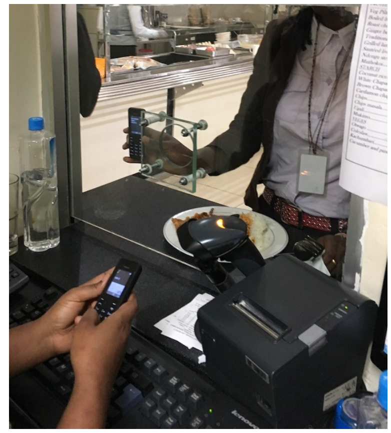


Figure 10: The cafeteria cashier (recipient) is operating the phone on the left, while a student participant (payer) is displaying $Code_2$ through the window to authorise the payment

Figure 11: Participants displaying the back of their phones (yellow DigiTally label shown on the back)

nod. When we mentioned this observation, most participants agreed that "*it's much easier*", and gave the example of the cafeteria. In general, once participants had some experience with the system, we noticed that they would always show the code and never speak it even in quiet areas. In the transaction flow people developed, visual cues were used to complete each major step, starting with the customer declaring that they want to use DigiTally by showing a yellow label on the back of the phone that identifies the participant by name (Figure 11). However, we should not neglect the fact that some users in the target population for financial inclusion might be visually impaired.

6.2 Error Recovery

We designed error recovery to prompt the user for three dummy inputs (any sequence of digits) to ensure that they were fully aware that they were indeed resetting the saved transaction data; in our implementation, a wrong reset could lead to inconsistent states on the SIMs involved. Inconsistent states currently require a 'hard reset' of the relationship between the two users by deleting the contact information and creating a new contact on each phone.

Before each prompt for $Code_2$, the recipient is shown a confirmation of the transaction details. After the dummy inputs are provided, the session is reset by discarding $Code_1$ as well as any intermediate results that were saved for validating $Code_2$. Now the recipient (merchant) can start a fresh transaction: the next time a transaction is initiated between the same two parties, a new $Code_1$ is generated. Even with default-text entry enabled in the applet to fill in the previously entered code, participants requested an easier way to rectify a mistake.

As error recovery was performed only by one student participant during the actual trial, most were likely evaluating the training they received rather than anything they actually did. Two users needed to reset a transaction during training. One participant commented: "*For the error issue: I think you should make it simpler to correct the error and provide instructions on how to correct it once the error is made since you will not be around to show them. For example, when an error of different amounts that needs one to put code 1 three times, I think you should write 'repeat code 1 three times to correct the error'.*" The other participant agreed: "*If there's a simpler way of resetting a wrongly transacted code DigiTally will be better.*" None of the other participants, including the merchants, had to use this error-recovery method, likely due to the training component instructing users to agree on the amount before proceeding with the transaction. We emphasised the importance of this step; perhaps the 'burden' of going through the error recovery process helped motivate careful exchange of the codes.

Our prototype error-recovery mechanism needs a redesign. A simple alternative would be a menu option to reset transaction data, requiring authentication with the user's PIN rather than requiring them to enter dummy inputs.

7. DISCUSSION

This paper reports the results of a preliminary study with an early prototype of DigiTally, tested on participants from a university in Nairobi. We described how DigiTally involves sharing two 8-digit codes to protect users from unauthorised payments while at the same time allowing them freedom from reliance on network coverage. This has an advantage that the payment service operator has no marginal transaction costs and can offer zero fees for some transactions. However, the code-sharing process has always presented a possible usability challenge. To test this and other aspects of this early prototype, DigiTally was introduced to Kenyan participants in order to identify and resolve any major usability issues before a later field trial with a more representative sample of service users.

The errors made in sharing codes suggest a need to make recovery from this type of error more intuitive. Nevertheless, most participants completed the task with few code-sharing errors. The average speed for every student was less than a minute, and overall the average transaction speed was close to half a minute. Given our observations of participants and the comments they made, it seems that the process, overall, seemed straightforward and fast. The SUS scale gives insight into the extent to which DigiTally's usability might inhibit its use. One of the lessons learned was the need to demonstrate clearly to users: (a) how errors can be avoided; and (b) how the codes prevent cheating.

Observations with initial trial users and open-ended answers following the trial indicate that users' main fear is the difficulty of recovery from errors. This is already an issue for M-Pesa when people send payments to the wrong phone number (it is a well-known problem, and one of the authors experienced it first-hand). Recoverability is especially important in the initial trials and adoption phase of a new system because expert help will not be available in remote villages; users' ability to figure out how to recover from mistakes may well be the difference between their adopting DigiTally and rejecting it.

Higher SUS scores tend to predict loyalty and word-of-mouth recommendations [18]. Users with scores over 80 (c.f. the average score of 83.1 for student participants) are called 'Promoters' because they are more likely to recommend a system, while users with a score below 60, called 'Detractors', are more likely to say negative things about a system. DigiTally is in a position to create 'Promoters', especially when it comes to student participants.

There are various things we can do to make DigiTally more usable for busy merchants. For example, we can give merchants a smartphone app that reads the customer's phone number from a QR code on a sticker on their phone, display $Code_1$, and read $Code_2$ from their phone screen. This way the merchant does no more work than with M-Pesa. Two of the merchants preferred DigiTally to M-Pesa because of speed; DigiTally does not force a cashier to wait for up to half a minute while a payment confirmation makes its way through Nairobi's congested mobile network. DigiTally takes more keystrokes, but the outcome is then immediate.

The fact that DigiTally has zero marginal costs also means that it can be offered as a zero-fee payment mechanism between friends and family, like personal cheques in the UK. Loan clubs, such as Rotating Savings and Credit Associations (ROSCAs) or savings clubs, and money-guards (used to enforce savings through the commitment of funds), whether formal or informal, are an important part of the financial ecosystem in many less developed countries. See Collins *et al.* [6] and Banerjee and Duflo [3] for more information about financial inclusion and the financial tools used by the poorest demographics (living on less than $2 a day).

Given our relatively high SUS, learnability, and usability scores, the challenge is encouraging first time use. The factor most apparently affecting first time use, based on free-text responses, was the 8-digit codes. The most frequent suggestion was to remove or reduce them. Although many participants were familiar with M-Pesa, and thus sharing an 11-digit phone number, this is potentially less time-consuming and error-prone because it is the same number every time for the recipient (their own phone number) and a single time entry for the sender. It also requires no input from the recipient, so a merchant can focus on other tasks.

Participants perceived DigiTally to be secure, and the codes were the reason behind this perception. As mentioned in section 5.3.1 item 3, participants asked if these codes can be shortened. However, this would not be possible without compromising the security of the system. We discuss the DigiTally protocol, the cryptography, and security parameters and features in the technical paper [4]. A few participants also requested a better money tracking tool: some participants liked that statistics about their transactions were available (e.g. amounts sent and received as explained in section 4.4), and some users requested the ability to view more than just the last transaction (as discussed in section 5.3.1 item 4). This is not a critical issue, as we can engineer the system to include more transactions as allowed by hardware constraints and storage capacity.

Because migrant remittances are a key application of mobile payment systems (such as M-Pesa), some of our participants discussed the possibilities of remote transactions (see section 5.3.1 item 4). We designed DigiTally primarily for face-to-face offline transactions and not for remote transactions that would require a medium to exchange the codes. However, if such a medium exists, then DigiTally codes can be exchanged online to process remote transactions. Codes can be exchanged using SMS or over the phone; they could even be exchanged by post. DigiTally can be viewed as a platform to do the cryptographic operations required to process payments, on users' phones rather than on a centralised server, and systems can be built on top of DigiTally to perform remote transactions – relying on the DigiTally SIM to do the cryptography.

DigiTally was praised for being quick, easy to use, and easy to learn. Kenyans are already familiar with M-Pesa, which DigiTally was designed to mimic; that may have helped directly, while other factors our subjects liked, such as being able to review transactions, may have been features that they would like to see in M-Pesa. Therefore the results in this paper do not measure how DigiTally would be perceived by users who have never used any mobile payment system.

But M-Pesa, like most mobile payment systems, cannot work offline, and therefore fails to provide service to the poorest communities. Just as M-Pesa appealed most strongly to people with phones but no payment cards, so also DigiTally should appeal most strongly to people with no network service at all.

This was confirmed when we made two field trips to scout possible sites for a second-round field trial. When we visited a small town near Nairobi with good network coverage, stakeholders were interested in playing with the system, but saw its main benefit as being potentially programmable so that it could support their specific applications. When we visited Busia, a rural community near Lake Victoria with very poor network coverage, stakeholders were delighted, and played with it for hours. They considered DigiTally to be just what they needed to solve their problems with network coverage and reduce transaction costs.

8. CONCLUSION

We designed and developed an early prototype of DigiTally, an offline phone payment system, and tested it on Kenyan participants in a preliminary study. In addition to error rates, transaction speed data, and SUS scores, we reported on supporting data from demonstration sessions (observed behaviours and comments) and free-text written responses at the end of the study. Our results indicate that participants found DigiTally easy to use and that they liked key aspects of the system, including its perceived security and that it did not require network coverage to process payments.

We have demonstrated that DigiTally can be used for making payments without network coverage or transaction fees. Some specific technical improvements are needed, most notably the process of recovery from errors. We discovered that while DigiTally is slightly less convenient to use than existing mobile payment systems, the added burden is not excessive; people in areas with poor network coverage are eager to use it. Furthermore, the burden mostly falls on merchants, and busy merchants are likely to have enough money to buy better terminals. There is a realistic prospect of developing DigiTally into a workable system that will extend mobile payments to the millions of people who are currently excluded from the world of electronic payments.

9. ACKNOWLEDGEMENT

This work was supported by a grant from the Bill & Melinda Gates Foundation, and approved by ethics committees at the University of Cambridge and at Strathmore University.

We thank the following colleagues at the Computer Laboratory, University of Cambridge: Laurent Simon for providing feedback throughout the project; Alastair Beresford, Kat Krol and Alexander Vetterl for helpful comments about the paper. We also thank colleagues at Strathmore University: Edwin Njeru, Nelson Mutua and David Mutyethau for their assistance with this project.

10. REFERENCES

[1] R. J. Anderson. UEPS–a second generation electronic wallet. In *European Symposium on Research in Computer Security*, pages 411–418. Springer, 1992.

[2] R. J. Anderson and S. J. Bezuidenhout. Cryptographic credit control in pre-payment metering systems. In *Security and Privacy*, page 15. IEEE, 1995.

[3] A. Banerjee and E. Duflo. *Poor economics: A radical rethinking of the way to fight global poverty.* PublicAffairs, 2012.

[4] K. Baqer, J. Bezuidenhoudt, R. Anderson, and M. Kuhn. SMAPs: Short Message Authentication Protocols. *International Workshop on Security Protocols*, 2016.

[5] G. Chen and S. Rasmussen. bKash Bangladesh: A fast start for mobile financial services. 2014.

[6] D. Collins, J. Morduch, S. Rutherford, and O. Ruthven. *Portfolios of the poor: how the world's poor live on $2 a day.* Princeton University Press, 2009.

[7] J. DeCoster. Scale construction notes. *Department of Psychology University of Alabama*, 2000.

[8] R. G. Downey and C. V. King. Missing data in Likert ratings: A comparison of replacement methods. *The Journal of general psychology*, 125(2):175–191, 1998.

[9] P. Dupas, S. Green, A. Keats, and J. Robinson. Challenges in banking the rural poor: Evidence from Kenya's western province. Technical report, National Bureau of Economic Research, 2012.

[10] L. Faulkner. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods*, 35(3):379–383, 2003.

[11] N. Hughes and S. Lonie. M-Pesa: mobile money for the "unbanked" turning cellphones into 24-hour tellers in Kenya. *Innovations*, 2(1-2):63–81, 2007.

[12] W. Jack and T. Suri. Mobile money: The economics of M-PESA. Technical report, National Bureau of Economic Research, 2011.

[13] R. Kainda, I. Flechais, and A. 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*, page 11. ACM, 2009.

[14] J. R. Lewis and J. Sauro. The factor structure of the system usability scale. In *International Conference on Human Centered Design*, pages 94–103. Springer, 2009.

[15] R. Macefield. How to specify the participant group size for usability studies: a practitioner's guide. *Journal of Usability Studies*, 5(1):34–45, 2009.

[16] J. Nielsen and T. K. Landauer. A mathematical model of the finding of usability problems. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 206–213. ACM, 1993.

[17] J. Sauro. Measuring usability. *A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices*, 2011.

[18] J. Sauro. How to measure learnability. `http://www.measuringu.com/blog/measure-learnability.php`, 2013. [Last accessed 01-November-2016].

[19] J. Sauro and J. R. Lewis. When designing usability questionnaires, does it hurt to be positive? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2215–2224. ACM, 2011.

[20] J. Sauro and J. R. Lewis. *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann, 2016.

[21] M. S. Uddin and A. Y. Akhi. E-wallet system for Bangladesh an electronic payment system. *International Journal of Modeling and Optimization*, 4(3):216, 2014.

[22] R. A. Virzi. Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 34(4):457–468, 1992.

[23] J. M. Zimmerman and S. Baur. Understanding how consumer risks in digital social payments can erode their financial inclusion potential. `https://www.cgap.org/sites/default/files/Brief-Understanding-How-Consumer%20Risks-in%20Digital-Social-Payments-March-2016.pdf`, 2016.

TurtleGuard: Helping Android Users Apply Contextual Privacy Preferences

Lynn Tsai[1], Primal Wijesekera[2], Joel Reardon[1], Irwin Reyes[3], Jung-Wei Chen[4], Nathan Good[4], Serge Egelman[1,3], and David Wagner[1]

[1]University of California, Berkeley, Berkeley, CA
{lynntsai,jreardon}@berkeley.edu, daw@cs.berkeley.edu

[2]University of British Columbia, Vancouver, BC
primal@ece.ubc.ca

[3]International Computer Science Institute, Berkeley, CA
{ioreyes,egelman}@icsi.berkeley.edu

[4]Good Research, Inc., El Cerrito, CA
{jennifer,nathan}@goodresearch.com

ABSTRACT

Current mobile platforms provide privacy management interfaces to regulate how applications access sensitive data. Prior research has shown how these interfaces are insufficient from a usability standpoint: they do not account for *context*. In allowing for more contextual decisions, machine-learning techniques have shown great promise for designing systems that automatically make privacy decisions on behalf of the user. However, if such decisions are made automatically, then feedback mechanisms are needed to empower users to both audit those decisions and correct any errors.

In this paper, we describe our user-centered approach towards designing a fully functional privacy feedback interface for the Android platform. We performed two large-scale user studies to research the usability of our design. Our second, 580-person validation study showed that users of our new interface were significantly more likely to both understand and control the selected set of circumstances under which applications could access sensitive data when compared to the default Android privacy settings interface.

1. INTRODUCTION

Smartphones store a great deal of personal information, such as the user's contacts, location, and call history. Mobile operating systems use *permission systems* to control access to this data and prevent potentially malicious third-party applications ("apps") from obtaining sensitive user data. Part of the purpose of these permission systems is to inform and empower users to make appropriate decisions about which apps have access to which pieces of personal information.

The popular open-source Android mobile platform has used two general approaches to give users control over permissions. Initially, permissions were presented as an install-time ultimatum, or ask-on-install (AOI): at installation, an application would disclose the full list of sensitive resources it wished to access. Users could either grant access to all requested permissions or abort the installation entirely. Prior research has shown that most users do not pay attention to or do not these prompts when shown at install-time [12].

Recently, an *ask-on-first-use* (AOFU) permission system replaced install-time disclosures on Android. Under AOFU, the user is prompted when an application requests a sensitive permission for the first time. The user's response to this permission request carries forward to all future requests by that *application* for that *permission*. The AOFU approach, however, fails to consider that the user's preferences may change in different contexts. It only learns the user's preferences once under a certain set of contextual circumstances: the first time an application tries to access a particular data type. This system does not account for the fact that subsequent requests may occur under different contextual circumstances and therefore may be deemed less appropriate. For instance, a user might feel comfortable with an application requesting location data to deliver desirable location-based functionality. The same user, however, might find it unacceptable for the same application to access location for the purposes of behavioral advertising, possibly when the application is not even being used.

The *contextual integrity* framework can explain why AOFU is insufficient: it fails to protect user privacy because it does not account for the context surrounding data flows [25]. That is, privacy violations occur when a data flow (e.g., an app's access to a sensitive resource) defies user expectations. In recent work [38, 39], we showed that mobile users *do* make contextual privacy decisions: decisions to allow or deny access are based on what they were doing on their mobile devices at the time that data was requested.

In theory, asking the user to make a decision for every request is optimal, as the user will be able to account for the surrounding context and can then make decisions on a case-by-case basis. In practice, however, this results in unusable privacy controls, as the frequency of these requests could overwhelm the user [38]. Consequently, automating these decisions with machine learning yields a balance between

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

accurately implementing users' privacy preferences and not overburdening them with too many decisions [39]. Such automation requires the platform to have feedback mechanisms so that automated decisions can be reviewed and errors can be corrected, thereby yielding fewer future errors.

To this end, we designed a novel permission manager, TurtleGuard, which helps users to vary their privacy preferences based on a few selected contextual circumstances. It also provides information about the apps that they use, by providing a feedback loop for them to audit and modify how automated decisions are made. TurtleGuard allows users to (i) vary their decisions based on the visibility of the requesting application – our previous work showed that the visibility of the requesting application is a critical factor used by users when making mobile app privacy decisions [38], and (ii) have an improved understanding of how third-party applications access resources in the real world and under varying contextual circumstances.

We conducted an initial 400-person experiment to evaluate our preliminary design. Based on our analysis of this data, we then iterated on our design, conducting a 580-person validation study to demonstrate our design's effectiveness. Both experiments had four tasks: three tasks that involved using the system to locate information about current application permissions, and one task that involved modifying settings. We observed that participants who used TurtleGuard were significantly more likely to vary their privacy preferences based on surrounding circumstances than the control group. We believe that these results are a critical contribution towards empowering mobile users to make privacy decisions on mobile phone platforms. Our contributions are as follows:

- We present the first contextually-aware permission manager for third-party applications in Android.
- We show that when using our new interface (compared to the existing Android interface) participants were *significantly* more likely to both understand when applications had foreground versus background access to sensitive data and how to correctly control it.
- We show that our proposed interface has a minimal learning curve. Participants, who had never used TurtleGuard before, were as successful at accomplishing information retrieval tasks as those who used the existing Android interface.

2. RELATED WORK

The Android OS has thus far used two different permission models: ask-on-install (AOI) permissions, and ask-on-first-use (AOFU) permissions. Versions of Android before version 6.0 (Marshmallow) implemented ask-on-install permissions. Under this model, applications request that the user grant all permissions to the application at install time. The user must consent to all requested permissions in order to complete installation. Otherwise, if the user wishes to deny any permission, the only option available is to abort the installation entirely. Research has shown that few users read install time permissions, and fewer still correctly understand their meaning [12, 18].

Versions of Android from 6.0 (Marshmallow) onward use the AOFU permission model instead. Under AOFU, applications prompt users for sensitive permissions at runtime. These prompts protect access to a set of 24 "dangerous permissions," including geolocation data, contact lists, and SMS. Prompts appear when the application attempts to request protected resources for the first time. This has the advantage of giving users contextual clues about why an application requires a protected resource: users can consider what they are doing when the prompt appears to help determine whether to approve the request. Although AOFU offers an improvement over the install-time model in this regard, first-use prompts insufficiently capture a user's privacy preferences [39]. That is, the AOFU model does not consider scenarios where an application requests access to data under varying contexts.

Research on permission models has found that users are often unaware how apps access protected resources and how access may be regulated [12, 8, 11, 36, 34]. Shih et al. showed that users are more likely to disclose privacy information when the purpose is unclear [35]. Prior work has specifically analyzed location data: Benisch et al. show that a vast number of factors (time, day, location) contribute to disclosure preferences [5]; Reilly et al. show that users want minimal interaction with their technology [31]. Additionally, Patil et al. takes into consideration context: they suggest making feedback actionable and allowing for selective control regarding location data [29]. They also show that users have difficulty articulating location access controls, and suggest an interface that includes contextual factors as a potential solution [28]. Almuhimedi et al. studied AppOps, a permission manager introduced in Android 4.3 but removed in Version 4.4.2 [1]. AppOps allowed users to review and modify application permissions once installed, as well as set default permissions that newly installed applications must follow. They examined privacy nudges that were designed to increase user awareness of privacy risks and facilitate the use of AppOps. They concluded that Android users benefit from the use of a permission manager, and that privacy nudges are an effective method of increasing user awareness [1].

Although AppOps was removed from Android, Android 6.0 (*Marshmallow*) reintroduced permission management. It—and subsequent versions as of this writing—include an updated interface that allows the user to view all of the permissions that a particular app has been granted, as well as all of the apps that have been granted a particular permission (Figure 1). Unfortunately, these controls are buried deep within the Settings app, and it is therefore unlikely that users are aware of them. For instance, viewing a particular app's permissions requires navigating four levels of sub-panels, whereas viewing all the apps that have requested a particular permission requires navigating five levels. By comparison, TurtleGuard is one click from the main Settings panel and explicitly presents the relationships between applications, permissions, and controls.

XPrivacy [6], DonkeyGuard [7], Permission Master [23], and LineageOS's[1] Privacy Guard [24] are examples of other third-party permission management software. These utilities require additional privileges and techniques to install because Android provides no official mechanisms for third-party programs to modify the permission system. For instance, Privacy Guard is built into the LineageOS custom ROM [24];

[1]LineageOS is a recent fork of CyanogenMod after the latter's discontinuation.

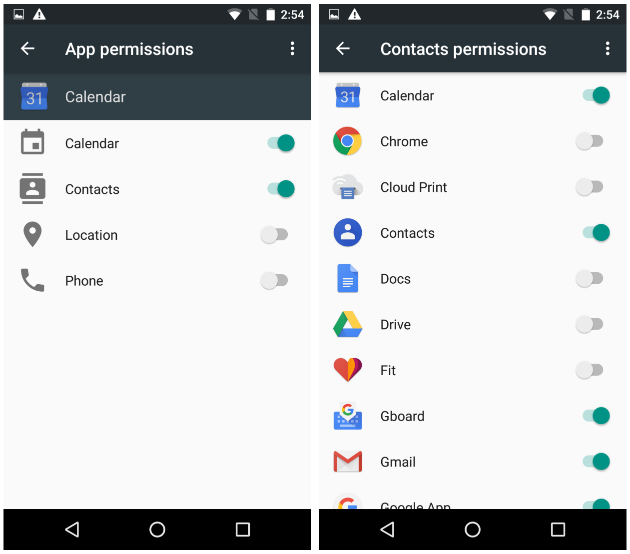


Figure 1: After navigating four and five levels of sub-panels within the Android Settings app, respectively, users can limit a specific app's access to specific permissions (left) or limit the apps that can access a particular permission (right).

others use the Xposed Framework [32], which requires an unlocked bootloader and a custom recovery partition. Such restrictions are necessary to prevent malicious software from interfering with the existing permission system.

Third-party permission managers offer users a variety of features to fine-tune access to sensitive resources on their devices. XPrivacy has the option to pass fake data to applications that have been denied access to protected resources [2]. Hornyack et al.'s AppFence similarly allows users to deny permissions to applications by providing fake data [16]. Providing fake data is more desirable than simply failing to provide any data at all, as the latter may cause functionality loss or application failures.

These managers follow an Identity Based Access Control model (IBAC), where individual permissions can be set for each app. Although this model allows users to specify fine-grained permission preferences, this may be ineffective in practice for two reasons. First, users may be overwhelmed by the number of settings available to them, some of which are only tangentially relevant to privacy. This security design failure is known as the *wall of checkboxes* [14]. XPrivacy and Permission Master show controls for resources whose direct effects on user privacy are unclear, such as keeping a device awake. TurtleGuard improves usability by showing only controls for resources deemed "dangerous" in the Android platform [15] and others that previous research has shown are conducive to using run-time prompts [10]. Second, none of the existing permission managers display the context in which protected resources were accessed. XPrivacy, Donkey Guard, and LineageOS's Privacy Guard provide timestamps for resource accesses, but the user does not receive important information about the app's state, such as whether it was actively being used when it requested access to sensitive data. Permission Master offers no historical information at all. TurtleGuard partially addresses this problem by listing recently allowed and denied permission access requests, along with the state and visibility of the requesting application at the time the permission was requested.

Apple's iOS platform offers visibility-sensitive location privacy settings: "Never" and "Always" (the two settings analogous to Android's permission on/off toggles), and a "While using the app" option, which only permits an application to access location data while the application is active on the screen. TurtleGuard uses the same options, but our design is novel in both the extent of these settings and in who controls them. Apple's iOS allows developers to control which of the three options are available to users to select [3]. Application developers have faced criticism for removing the "While using the app" option, forcing users to choose between reduced functionality and granting the application unrestricted access to sensitive location data [26]. Our design, by contrast, gives users all three of these options for all *sensitive* permissions (Table 5, Appendix). Furthermore, developers cannot restrict user choice with these settings, as TurtleGuard is implemented in the operating system.

Wijesekera et al. show that although AOFU improves on install-time permissions, AOFU is insufficient because it does not account for the context of the requests [39]. They examined this by instrumenting the Android platform to log all instances of apps accessing sensitive resources. In addition to their instrumentation, the platform randomly prompted users about the appropriateness of various permission requests as those requests occurred. Participant responses to these prompts were treated as the dependent variable for a training set. Their study showed that 95% of participants would have chosen to block at least one access request had the system notified them. On average, participants would have preferred to block 60% of permission requests. Indeed, other work suggests that contextual cues are key in detecting privacy violations [25, 4].

A natural extension of AOFU is "ask on *every* use": rather than extrapolating the user's first-time preference to all future accesses to a given resource, each access instead requires user input. Such a model would conceivably allow users to accurately specify their contextual preferences because they know exactly which app attempted to gain access to what resource under which circumstance. This approach, however, is unusable in practice. Research has shown that applications request access to permission-protected resources with great frequency: on an average smartphone, roughly once every 15 seconds [38]. Such a high frequency not only risks habituation, but would render the device inoperable.

Recent research on permission models has turned towards using machine learning (ML) [39, 20, 21, 19]. One advantage is ML's ability to incorporate nuanced contextual data to predict user preferences; the approach has shown significantly lower error rates over the *status quo*, i.e., AOFU. Wijesekera et al. [39] also showed that ML reduces user involvement, thereby minimizing habituation. They emphasize, however, the importance of having a user interface that functions as a feedback-loop to the classifier, since no practical classifier will ever be 100% accurate. Users can use the interface to audit the decisions made by the classifier and correct any decisions that do not match their preferences.

Such a mechanism not only ensures that the classifier improves its accuracy over time, it also keeps users aware of decisions that were made on their behalves and informs them of how third-party apps are accessing sensitive resources under various circumstances.

TurtleGuard provides two core components necessary for usability under such contextual privacy models: we provide users with key contextual information when regulating access to sensitive resources, and we provide a method for users to audit and correct the decisions that have been automatically made by the system.

3. DESIGN OBJECTIVES

TurtleGuard's primary function is to inform users about the decisions that have been automatically made on their behalf, while allowing them to easily correct errors (thereby improving the accuracy of future decisions). These errors can be either false positives—an app is denied a permission that it actually needs to function—or false negatives—an app is granted access to data against the user's preferences.

Thompson et al. showed how attribution mechanisms can help users better understand smartphone application resource accesses [37]. They found that users expect this information to be found in the device's *Settings* app. In our initial experiment, we evaluated TurtleGuard as a standalone app, though for this reason we ultimately moved it within the Android *Settings* panel prior to our validation experiment.

3.1 Incorporating Context

In prior work, researchers observed that only 22% of participants understood that applications continue to run when not visible and that they have the same access to sensitive user data that they do when actively being used [37]. This means that the majority of users incorrectly believe that applications either stop running when in the background or lose the ability to access sensitive data altogether. Wijesekera et al. corroborated this observation in a field study of users' privacy expectations: users are more likely to deem permission requests from background applications as being inappropriate or unexpected, and indicate a desire to regulate applications' access to sensitive data based on whether or not those applications are in use [38].

In the default permission manager, users cannot vary their decisions based on the visibility of the requesting application, or any other contextual factors. Our overarching goal is to empower users to make contextual decisions (i.e., based on what they were doing on the device) and to apply these decisions to future use cases, so that fewer decisions need to be explicitly made overall. As a first step towards allowing users to make contextual decisions, TurtleGuard makes decisions about whether or not to allow or deny access based on whether the requesting application is actively being used. While this is but one contextual factor amongst many, it is likely one of the most important factors [38].

Moving one step beyond the all-or-nothing approach of allowing or denying an application's access to a particular data type, our new design gives the user a third option: allowing applications to access protected data only *when in use* (Table 1 and Figure 2). When the *when in use* option is selected, the platform only allows an application to access a resource if the application is running in such a way that it is *conspicuous* to the user of the device. We consider the following behaviors conspicuous: (i) the application is running in the foreground (i.e., the user is actively using it), (ii) the application has a notification on the screen, (iii) the application is in the background but is producing audio while the device is unmuted. If these conditions do not hold, then access to the resource is denied.

option	meaning
always	The permission is always granted to the requesting application, regardless of whether the application is running in the foreground or background.
when in use	The permission is granted to the requesting application only when there are cues that the application is running, and denied when the application is running invisibly in the background.
never	The permission is never granted to the requesting application.

Table 1: The three possible permission settings under TurtleGuard. The *when in use* option accounts for the visibility of the requesting app, which is a strong contextual cue.

3.2 Auditing Automatic Decisions

Although Android currently provides an interface to list the applications that recently accessed location data, similar information is unavailable for other protected resources. The existing Android interface also does not differentiate between actions that applications take when *in use* and when *not in use*. TurtleGuard's main design objective is therefore to communicate the types of sensitive data that have been accessed by applications and under what circumstances.

Our initial design of TurtleGuard can be seen in Figure 2. The first tab (ACTIVITY) shows all of the recently allowed or denied permission requests, including when those requests occurred and whether the application was in use at the time. TurtleGuard presents this information as a running timeline—a log sorted chronologically. A second tab lists all of the apps installed on the phone in alphabetical order, allowing the user to examine what decisions have been made for all permissions requested by a particular app. The user can expand a log entry to change future behavior, if the platform's automated decision to allow or deny a permission did not align with the user's preferences. When the user uses this interface to change a setting, the classifier is retrained based on the updated information.

3.3 Permission Families

Android uses over 100 permissions and a given resource can have more than one related permission. Felt et al. found that not all the permission types warrant a runtime prompt—it depends on the nature of the resource and the severity of the threat [9]. Consequently, TurtleGuard only manages a subset of permissions (Table 5, Appendix) based on those deemed sensitive by prior work and by the latest Android version. In the first prototype of TurtleGuard, we had listed the original names of the permissions, ungrouped. One of the changes we made as we iterated on our design after our pilot experiment was to implement permission "fami-

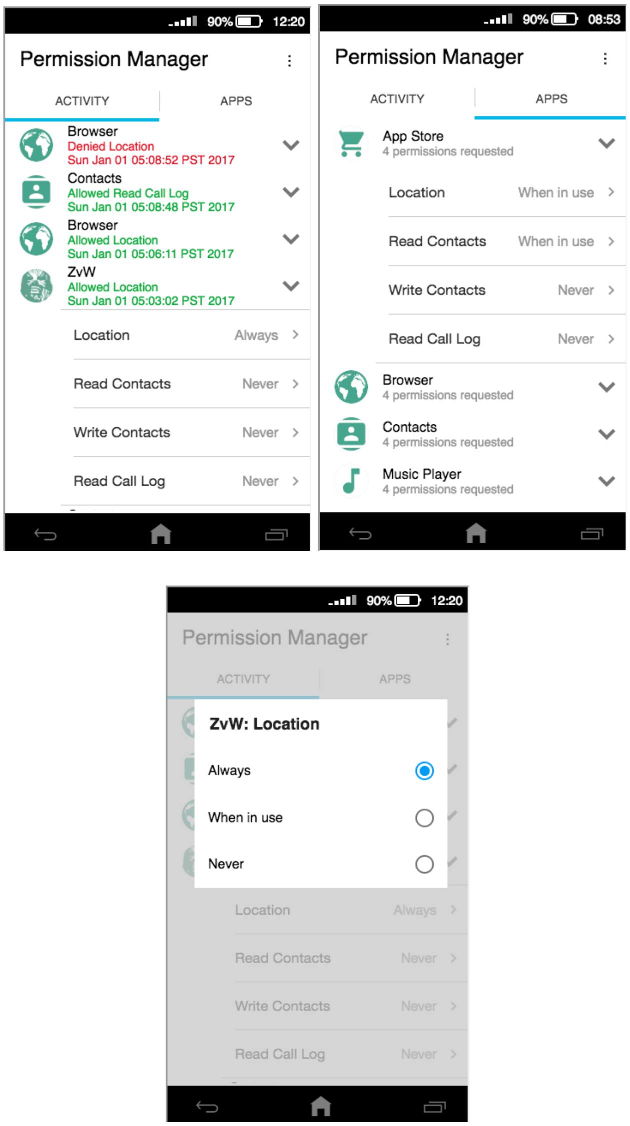


Figure 2: The pilot design of TurtleGuard listed recent app activity (top left), a list of installed apps and their associated permissions (top right). Permissions can be *always* granted, granted only *when in use*, or *never* granted (bottom).

lies." For example, READ_CONTACTS and WRITE_CONTACTS are grouped into a single CONTACTS permission family. This means that within TurtleGuard, users only see the human-readable resource type and not the underlying permissions the family manages. Any changes that a user makes about granting a resource therefore affects all permissions in the same family. For example, there is no longer a distinction between coarse and fine location data; both are either allowed or denied by a location settings change made using the TurtleGuard interface.

4. METHODOLOGY

We conducted two online experiments to evaluate the effectiveness of TurtleGuard at providing users with information and control over app permissions, as compared to Android's default permission manager (as of versions 6.0). We designed the first experiment to examine our initial prototype, as described in the previous section. Based on the analysis of our first experiment, we made changes to our design, and then validated those changes through a second experiment. In both experiments, we asked participants to perform four different tasks using an interactive Android simulation. These tasks involved either retrieving information about an application's prior access to sensitive resources or preventing access in the future (i.e., modifying settings). Our study was approved by our IRB (#2013-02-4992).

In both experiments, we randomly assigned participants to either the *control* or *experimental* conditions. We presented *control* participants with an interactive HTML5 simulation of the default permission manager, which is accessible from within the Settings app. We presented *experimental* participants with an interactive HTML5 simulation of our novel permission manager, TurtleGuard. During our pilot experiment, TurtleGuard was accessible through an icon on the home screen labeled "Privacy Manager," though we added it as a sub-panel to the Settings app prior to the validation experiment (Figure 6 in the Appendix). The questions and tasks for participants were identical for the two conditions and both experiments.

4.1 Tasks

We presented participants with four tasks to complete using the interactive Android simulations: three tasks to retrieve information about permission settings, and one task to modify permission settings. Some of these tasks required participants to find information about a specific app's abilities. In order to avoid biases from participants' prior experiences and knowledge of specific real-world apps, these questions instead focused on a fictitious app, *ZvW*. While we randomized the order of the tasks, we ensured that Task 3 always came before Task 4 (i.e., we never asked them to prevent background location data collection prior to asking them whether background location data was even possible). After each task, we asked participants to rate the difficulty of the task using a 5-point Likert scale ("very easy" to "very difficult"). Finally, upon completing all tasks, we asked them several demographic questions and then compensated them $2. We now describe the four tasks in detail.

Task 1: What were the two most recent applications that accessed this device's location?

In this task, we asked participants to use the Android simulation and identify the two applications that most-recently accessed location data. Participants used two open-ended fields to answer this question. In the *control* condition, this task was correctly accomplished by navigating to the "location" screen from within the Settings application (Figure 3). This screen presents information about applications that recently requested location data.

In the *experimental* condition, this task was accomplished by simply studying the "activity" screen, which was displayed immediately upon opening TurtleGuard (Figure 2). Given that this task was already supported by the default permission manager, we wanted to verify that TurtleGuard performed at least as well.

Task 2: Currently, which of the following data types can be accessed by the ZvW application?
In the *control* condition, this was accomplished by performing the four steps to access the screen in Figure 4 (right): selecting the "Apps" panel within the Settings app (Figure 3, left), selecting the ZvW application, and then selecting the "Permissions." This screen depicted a list of permissions available to the application based on what the application declares as its required permissions; the user is able to fine-tune this by selectively disabling certain permissions using the sliders on this screen. We wanted participants to identify the permissions that were enabled, rather than all of those that *could* be enabled in the future.

In the *experimental* condition, participants could accomplish this task by selecting the "Apps" tab from within TurtleGuard and then expanding the ZvW application to view its requested permissions (Figure 2, top right). In both conditions, the correct answer to the question was that "location" is the only data type that can be accessed by the ZvW application. Again, given that this task was already supported by the default permission manager, we wanted to verify that TurtleGuard performed at least as well.

Task 3: Is the ZvW application able to access location data when it is not being actively used?
We designed this task to determine whether TurtleGuard was effective at communicating to participants in the *experimental* condition the difference between foreground and background data access. Similarly, we wanted to examine whether participants in the *control* condition understood that once granted a permission, an application may access data while not in use. Based on the settings of the simulations, the correct answer across both conditions was "yes."

Participants in the *control* group must navigate to Settings, then the "Apps" panel, and view the list of permissions corresponding to the ZvW application, similar to Task 2. Location is turned on, and so participants must be able to understand that this means that the permission is granted even when it is not actively being used. Participants in the *experimental* condition can use TurtleGuard's "Apps" tab to view the requested permissions for the ZvW application. This shows that the location permission is listed as "always" (Figure 2, top right) and that "when in use" is an unselected option (Figure 2, bottom).

Task 4: Using the simulation, prevent ZvW from being able to access your location when you aren't actively using ZvW (i.e., it can still access location data when it is being used). Please describe the steps you took to accomplish this below, or explain whether you believe this is even possible.
As a follow-up to the third task, the fourth task involved participants explaining the steps that they went through in order to limit background location access, or to explain that it is not possible.

Those in the *experimental* condition could locate and change this permission setting either through the activity timeline or by locating ZvW from the "Apps" tab (Figure 2). We marked answers correct that specifically mentioned changing the setting to "when in use."

Those in the *control* condition could not prevent this access. We marked responses correct if they indicated that this task was impossible to complete. Two coders independently reviewed the responses to this task (Cohen's $\kappa = 0.903$). The objective of this task was to see TurtleGuard's success at allowing participants to vary settings based on application use (a strong contextual cue) and to examine whether participants knew that this was not possible when using the default permission manager.

4.2 UI Instrumentation
We built an interactive HTML5 simulation of the UI designs described in the previous section using *proto.io*. We instrumented the simulations to log all interactions (e.g., panels visited, buttons clicked, etc.). This data allowed us to analyze how participants navigated the UI to perform each task.

4.3 Qualitative Data
In addition to analyzing the participants' responses to the four tasks, their perceived difficulty of each of the tasks, and their demographic information, we also collected responses to two open-ended questions:

Thinking about the tasks that you performed in this survey, have you ever wanted to find similar information about the apps running on your smartphone?
We coded participants' responses as a binary value. Responses indicating sentiments such as "yes" and "I always wanted that" were coded as true. Clear negative answers and weak affirmative answers such as "sometimes" and "maybe" were coded as false. The purpose of this question is to see how prevalent seeking information is in the real world.

Thinking about the simulation that you just used, what could be done to make it easier to find information about how apps access sensitive information?
We coded participants' responses in multiple ways. First, as binary values indicating contentment with the presented design. Responses that affirmed that the user would change nothing about the presented design were coded as true. Any complaints or suggestions were coded as false, as well as responses with uncertainty, confusion, or ambivalence (e.g., "I don't know"). We further coded responses that had specific suggestions, using tags for the different themes.

Each response was coded by two experienced coders working independently, who then compared responses and recorded their coding conflicts. The coders discussed and reconciled the differences using their mutually agreed upon *stricter* interpretation given the nature of the tasks. This produced the final coding of the data, which is used in our analysis.

5. PILOT EXPERIMENT
Using the methodology outlined in the previous section, we recruited 400 participants from Amazon's Mechanical Turk for a pilot experiment. We discarded 8 incomplete sets of responses, leaving us with 392 participants. Our sample was biased towards male respondents (65% of 392), however, a chi-square test indicated no significant differences between genders with regard to successfully completing each task. Disclosed ages ranged from 19 to 69, with an average age of 33. In the remainder of this section, we describe our results for each task, and then describe changes we made to

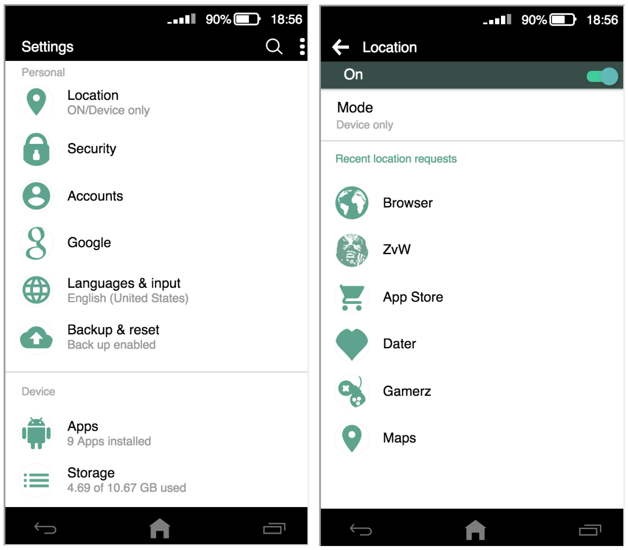


Figure 3: In Task 1, participants in the *control* condition could identify the most recent applications that requested location data from within the Settings application. This was also a valid method for Task 1 in the *experimental* condition for the validation study.

TurtleGuard's interface as a result of this initial experiment. We note that in our simulation, *Settings* can only be reached by tapping on the icon from the home screen. In all of our tasks, we also asked participants to evaluate perceived difficulty using a 5-point Likert scale.

5.1 Task 1: Recent Location Access

In the *control* condition, 84% of participants (167 out of 198) correctly completed this task, whereas only 68% (132 out of 194) completed it correctly in the *experimental* condition. This difference was statistically significant ($\chi^2 = 14.391$, $p < 0.0005$), though with a small-to-medium effect size ($\phi = 0.192$). In both cases, answers were marked correct if they mentioned both the Browser and ZvW applications (Table 2). Of the 49 participants in the *experimental* group who tried but failed, 13 never opened TurtleGuard, and over 73% (36 of 49) entered "Browser" and "Contacts", which were the first two applications listed in the activity tab of the Permission Manager. The activity tab showed recent resource accesses in a chronological order—"Browser" had been denied a *location* request and "Contact" had successfully accessed *call logs*.

Participants did not seem to understand that the activity log presented entries related to *all* sensitive data types, not just location data. This confusion might also stem from their familiarity with the location access panel in stock Android, in which location access requests are presented in chronological order. We hypothesize that this confusion is addressable by redesigning the activity log to better distinguish between data types and allowed-versus-denied permission requests. One possible way of implementing this is to create separate tabs for allowed and denied requests, as well as to group similar data types together (rather than presenting all permission request activity in chronological order).

Condition	Correct	Incorrect
Task 1		
control	167 (84%)	31 (15%)
experimental	132 (68%)	62 (32%)
Task 2		
control	140 (70%)	58 (29%)
experimental	116 (59%)	78 (40%)
Task 3		
control	86 (43%)	112 (56%)
experimental	153 (78%)	41 (21%)
Task 4		
control	47 (23%)	151 (76%)
experimental	144 (75%)	49 (25%)

Table 2: Participants in each condition who performed each task correctly during the pilot experiment.

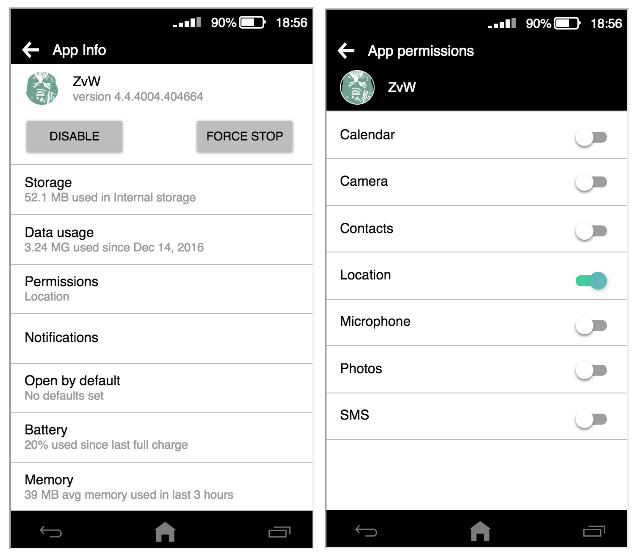


Figure 4: In Task 2, participants in the *control* condition could identify the permissions granted to the ZvW application by selecting the "Apps" panel from within the Settings application, and then selecting the application, followed by the "Permissions" panel.

5.2 Task 2: Finding Granted Permissions

In the second task, we asked participants to list all of the data types that the ZvW application *currently* had access to. We observed that 140 participants in the *control* condition (70.7% of 198) and 116 participants in the *experimental* condition (59.8% of 194) performed this task correctly. After correcting for multiple testing, this difference was not statistically significant ($\chi^2 = 5.151$, $p < 0.023$). However, despite the lack of statistical significance, we were surprised that not more people in the *experimental* condition answered correctly. Upon investigating further, we noticed several confounding factors that might have made this task more difficult for people in this condition. First, while the *control* condition displays the currently-allowed permissions as grayed-out text on the "App Info" page (Figure 4), the *experimental* condition lists all *requested* permissions—

which is a superset of the allowed permissions (top-right of Figure 2). Second, we noticed that due to an experimental design error, the permissions requested by the ZvW app in the *experimental* condition included several that were not included in the options presented to participants (e.g., "Write Contacts" and "Read Call Log"). This may have made this task confusing for these participants.

5.3 Task 3: Finding Background Activity

In the third task, we asked participants whether the ZvW application had the ability to access location data while not actively being used. We observed that 86 participants in the *control* condition (43% of 198) correctly answered this question, as compared to 153 participants in the *experimental* condition (78% of 194). This difference was statistically significant ($\chi^2 = 51.695$, $p < 0.0005$) with a medium effect size ($\phi = 0.363$). Thus, the new dashboard interface successfully differentiated between foreground and background permission usage.

5.4 Task 4: Limiting Background Activity

We observed that only 47 participants in the control condition (23% of 198) correctly stated that this task was impossible. In the *experimental* condition, 144 (74% of 193)[2] clearly articulated the steps that they would go through using the privacy dashboard to change location access from "always" to "when in use." This difference was statistically significant ($\chi^2 = 101.234$, $p < 0.0005$) with a large effect size ($\phi = 0.509$).

5.5 Design Changes

Based on the results of our first two tasks, in which participants in the *control* condition were more likely to correctly locate information about recent app activities and the permissions that apps had requested, we made several design changes to the TurtleGuard interface. First, we split the activity timeline into two separate tabs: recently allowed permission requests, and recently denied permission requests. Second, rather than showing all activity in chronological order, the activity timeline is now categorized by resource type, with the events for each resource type sorted chronologically. These changes can be seen in the top of Figure 5.

In addition to these changes, we also modified the apps tab to show grayed-out allowed permissions for each app, similar to the App Info panel in the default permission manager. Due to the error we noted in the *experimental* condition in Task 2, we made sure that all app permissions were the same in both conditions.

Finally, we moved TurtleGuard to be within the Settings app, so that it appears as a panel labeled "Permissions Manager" (Figure 6, Appendix). For consistency, when participants in the *experimental* condition select the "Permissions" sub-panel from within the "App Info" panel (Figure 4, left), they are now redirected to TurtleGuard's "Apps" panel, pre-opened to the app in question (Figure 5, bottom right).

6. VALIDATION EXPERIMENT

Following our pilot experiment and subsequent design changes, we performed a validation experiment. In the remainder of this section, we discuss our results (Table 3).

[2] One person could not load the `iframe` containing the simulation during this task.

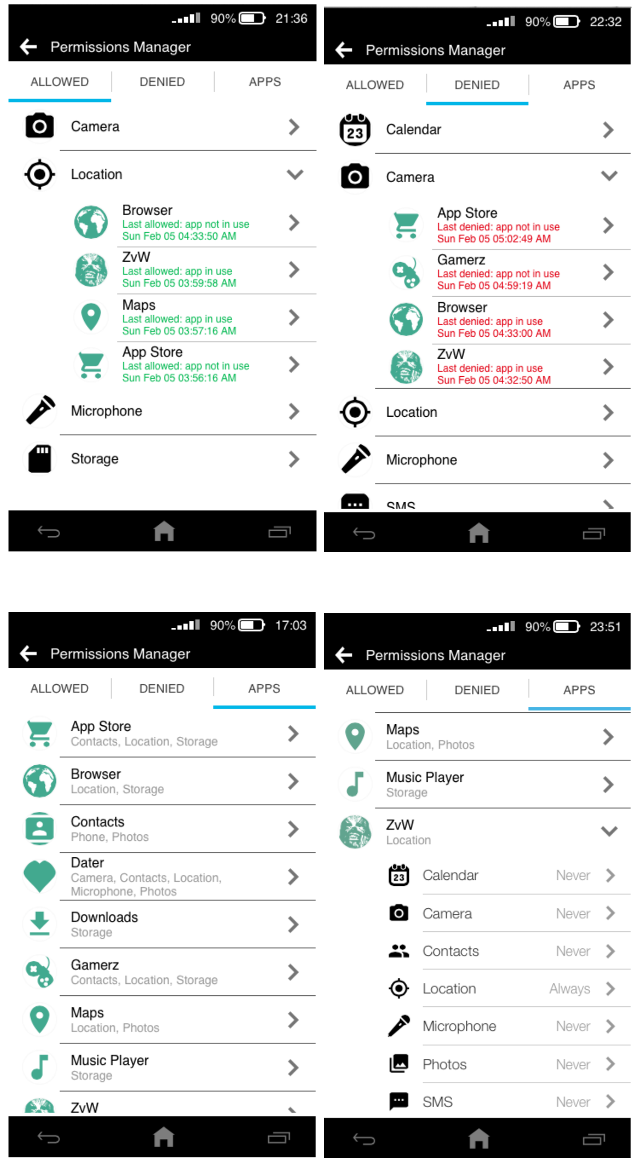


Figure 5: TurtleGuard separates recently allowed (top left) and denied (top right) permissions. The "Apps" tab lists the allowed permissions of all apps (bottom left). Expanding an app allows the user to make changes (bottom right).

6.1 Participants

Because of several known biases in Mechanical Turk's demographics [27, 33, 22], we decided to compare a sample of 298 Mechanical Turk participants to a sample of 300 Prolific Academic participants. Peer et al. recently performed several studies on various crowdsourcing platforms and concluded that the latter yields more diverse participants [30]. We limited both groups to participants based in the U.S., over 18, owning an Android phone, and having a 95% approval rating on their respective platform. After removing 18 incomplete responses, we were left with a combined sample of 580 participants. We analyzed the results from the two groups, and discovered that the high-level findings (i.e.,

Condition	Correct	Incorrect
Task 1		
control	237 (82.6%)	50 (17.4%)
experimental	241 (82.5%)	52 (17.5%)
Task 2		
control	232 (77.1%)	55 (22.9%)
experimental	226 (80.8%)	67 (19.2%)
Task 3		
control	108 (37.6%)	179 (62.4%)
experimental	230 (78.5%)	63 (21.5%)
Task 4		
control	79 (27.5%)	208 (72.5%)
experimental	224 (76.5%)	69 (23.5%)

Table 3: Participants in each condition who performed each task correctly during the validation experiment.

task performance) did not observably differ. For the remainder of our study, we therefore discuss the combined results. Our sample was biased towards male respondents (63% of 580), however, a chi-square test indicated no significant differences between genders with regard to successfully completing each task. Disclosed ages ranged from 19 to 74, with an average age of 33. Participants performed the same tasks as those in the pilot experiment and took on average 9 minutes and 17 seconds to complete the experiment.

6.2 Task 1: Recent Location Access

Recall that in this task, we asked participants to identify the two most recent applications that accessed location data. For the *experimental* condition, in addition to using the same method as the *control* (navigating to the "Location" sub-panel of the Settings app), participants could navigate to the "Allowed" tab within TurtleGuard, and then examine the "Location" permission to see the two most recent accesses (top left of Figure 5). In the *control* condition, 237 participants (82.6% of 287) correctly completed this task, whereas 241 (82.5% of 293) completed it correctly in the *experimental* condition. A chi-square test revealed that this difference was not statistically significant ($\chi^2 = 0.011$, $p < 0.918$).

We observed that most of the participants in both conditions used the default method of accomplishing this task (i.e., accessing the Location sub-panel): 80.1% of those who answered correctly in the *experimental* condition and 92.8% of those in the *control* condition. Fifteen participants in the *control* condition answered correctly despite not accessing the panel—likely by random guessing, and two who answered correctly by exhaustively searching the "App Info" panels of installed apps, to see which had been granted the location permission; 48 participants in the *experimental* condition used TurtleGuard to yield the correct answer.

A total of 102 participants incorrectly answered the question in Task 1. Of the incorrect responses, five participants failed to properly navigate the simulation and wrote that it was broken or the buttons did not work; 9 participants did not respond or wrote that they did not know. Of the other 88 participants, 38 (43%) listed "App Store" as one of their selections, making it the most common error.

More specifically, 18 participants listed their answers as both "App Store" and "Browser." We believe that this is because both the stock Android Apps Manager and TurtleGuard's "Apps" tab (Figure 5, bottom) sort the entries alphabetically, and by looking at the permissions available to both of these apps, participants would see that both have location access. Nevertheless, they are not the most *recent* apps to access location data.

Overall, these results suggest that the changes we made after our pilot experiment resulted in marked improvements. We further investigated this by examining participants' perceived ease-of-use, as measured using the 5-point Likert scale ("very easy (1)" to "very difficult (5)"). In the *experimental* condition, 84 participants accessed TurtleGuard to complete this task (regardless of whether or not they answered correctly). We compared these 84 responses with the 463 responses from participants who only used the default Settings panel (i.e., 195 in the *experimental* group and 268 in the *control* group). The median responses from both groups was "easy" (2), however there was a statistically significant difference between the groups (Wilcoxon Rank-Sum test: $Z = -3.9605$, $p < 0.0005$), with a small effect size ($r = 0.17$)—participants who used TurtleGuard found it more difficult compared to the default Settings panel. This difference appears to be due to those who performed the task incorrectly: the median response for TurtleGuard users who answered incorrectly was "difficult (4)," whereas it was "neutral (3)" for other participants. This may actually be a good thing: participants who confidently answered incorrectly are at greater risk due to over confidence, whereas those who had difficulty may be more likely to seek out more information.

6.3 Task 2: Finding Granted Permissions

In this task, participants had to locate the app's allowed permissions to discover that "location" was the only allowed permission in both the *experimental* and *control* conditions. This could be accomplished by viewing TurtleGuard's Apps tab (bottom of Figure 5) for those in the *experimental* condition, or by viewing an app's App Info panel from within the Settings app (Figure 4), which was available to those in either condition.

In total, 458 participants correctly performed this task (79% of 580). Table 3 displays the breakdown of the results by condition. A chi-square test did not yield statistically significant results between the two conditions in terms of task completion ($\chi^2 = 0.984$, $p < 0.321$).

Of the 226 *experimental* condition participants who performed the task correctly, 127 (56.2%) did so by using TurtleGuard. In total, 145 *experimental* condition participants accessed TurtleGuard, and reported a median task difficulty of "easy (2)." This did not significantly differ from the 375 other participants in both conditions who only examined the default Settings panels to perform the task and also reported a median difficulty of "easy" ($Z = 1.808$, $p < 0.238$); 60 participants never opened Settings (10 of whom answered the question correctly, likely due to random guessing).

6.4 Task 3: Finding Background Activity

To perform this task, participants in the *control* group had to navigate to Settings, then the "Apps" panel, and view the list of permissions corresponding to the ZvW application

(Figure 4). However, performing this sequence of steps still did not guarantee they would answer the question correctly: they needed to observe that location data was allowed, as well as understand that this meant that location data could be accessed by the app even when it is not actively being used. Participants in the experimental condition answered this question through TurtleGuard, which shows that the location permission was listed as "Always" (Figure 5), thereby eliminating the ambiguity.

We observed that 230 *experimental* condition participants answered this question correctly (78.5% of 293), as compared to only 108 *control* participants (37.6% of 287). A chi-square test showed that this difference was significant ($\chi = 97.914$, $p < 0.0005$) with a medium-to-large effect size ($\phi = 0.414$). This observation corroborates Thompson et al.'s findings [37] that users are largely unaware that apps can access sensitive data when not in use. TurtleGuard, however, was more effective at communicating this information to participants. Among the participants in the *experimental* condition, 24.57% took the extra step to click on the location entry (bottom right of Figure 5) to see the other options available (Figure 2): *always*, *when in use*, and *never*.

We found that 129 participants used TurtleGuard to perform this task, which suggests that 101 (34.5% of *experimental* condition participants) still got it correct either based on prior knowledge—a proportion consistent with Thompson et al.'s findings [37]—or after having used TurtleGuard in preceding tasks. There were 383 participants who completed the task by examining existing areas of the Settings app, whereas 68 participants never bothered to open Settings to complete this task. The median ease of use for those who used TurtleGuard was "easy (2)", while the median ease of use for those who used the default permission manager was "neutral (3)". This difference was statistically significant ($Z = -2.885$, $p < 0.004$) with a small effect size ($r = 0.13$). Participants in the *control* condition also took significantly longer to complete the task: 49.63 seconds versus 26.65 seconds. A Wilcoxon Rank-Sum test found this difference to be statistically significant ($Z = -5.239$, $p < 0.0005$, $r = 0.22$).

6.5 Task 4: Limiting Background Activity

Task 4 asked participants to describe the steps to prevent an application from accessing location data while the application was not in use, or to state that it is not possible to prevent it. It is only possible to prevent it using TurtleGuard.

In the *experimental* condition, 224 (76.5% of 293) explicitly stated how they would use TurtleGuard to change the permission to "when in use",[3] whereas only 79 (27.5% of 287) *control* group participants correctly stated that this task was impossible using the default permission manager. This difference was statistically significant ($\chi^2 = 137.14$, $p < 0.0005$) with a large effect size ($\phi = 0.49$).

A majority of the participants (72.5%) in the *control* group incorrectly believed that they could vary their decisions based on the *visibility* of the application. The most common responses involved disabling location data altogether, preventing the app from running, or restricting "background data":

- Settings > Apps > ZvW > Toggle Location Off
- Disable or Force Stop the Application
- Settings > Location > ZvW > Permissions > Toggle Location Off
- Settings > Apps > ZvW > Data Usage > Restrict Background Data
- Settings > Location > Toggle Location Off

A considerable portion (14%) chose to "restrict background data," which does something else entirely: it prevents data surcharges while roaming on foreign networks. This is another example of a disconnect between users' mental models and the true meaning of these configuration options. That said, a small number of participants in the *control* condition correctly stated that they would need to disable the app's location permission, and then re-enable it every time they wanted to use that app, a tedious process that is prone to forgetfulness—we treated this response as correct. Another substantial portion among the default permission manager condition (46%) wanted to block the location globally (from the default location panel) or block the location access from ZvW app entirely. While this is an overly restrictive option compared to *when in use*, this is the closest option provided in Android—we treated this as an incorrect response.

As expected, participants in the *control* condition rated the difficulty of this task as "neutral (3)", whereas the median Likert score from those in the *experimental* condition was "easy (2)". This difference was statistically significant with a large effect size ($p < 0.0005$, $\phi = 0.49$). The participants in the *control* condition who successfully completed the task (e.g., by acknowledging it was impossible) struggled immensely with it, evaluating it as "difficult (4)".

7. USER PERCEPTIONS

After completing the four tasks, participants answered two open-ended questions about whether they have looked for this type of permission information in the past, and whether they have any suggestions to offer us about the design of the interface they had just used. Two researchers independently coded each question and then resolved conflicts. We provide Cohen's inter-rater reliability statistic (κ) for each coding.

7.1 Prior Experiences

Our first question asked: *Thinking about the tasks that you performed in this survey, have you ever wanted to find similar information about the apps running on your smartphone?*

Our goal was to determine whether participants had previously thought about resource access or configuring privacy preferences, and whether having these features would be beneficial. On average, 63.1% of participants stated that they had thought about this (Cohen's $\kappa = 0.792$), and the experimental condition they were in proved to be insignificant. We did, however, observe a positive correlation between performance on the four tasks (i.e., number of tasks performed correctly) and reporting having previously thought about these issues ($p < 0.007511$, $r = 0.155$).

Among the people who chose to be more detailed in their responses, several themes emerged. A large portion mentioned that the reason they had tried these tasks before is that they wanted to be able to exert more control over their installed apps:

[3] We used a very conservative rubric: 11 participants who described using TurtleGuard, but did not explicitly use the phrase "when in use," were coded as being incorrect.

	Changes	No Changes
control	245 (85.4%)	42 (14.6%)
experimental	187 (63.8%)	106 (36.3%)

Table 4: Whether participants believed changes were needed to the interfaces they used during the validation study.

- "*I was somewhat familiar with these menus already before starting this task. I like to have control over my app permissions including location and data management.*"
- "*Yes, I've often wanted a little more control over what my apps get to access*"

A minority of participants expressed their frustrations on how the current default user interfaces in Android were confusing and did not let them set privacy preferences the way they wanted:

- "*Yes but usually can't find anything on there either like these. So I gave up trying.*"
- "*Yes. I want to know what they collect, although it gets tedious to try to figure it all out. Sometimes I'd rather just ignore it.*"

These comments highlight the fact that many users want to have control over resource usage by applications, and that many feel helpless to do so, given the options offered by current privacy management interfaces. These observations further strengthen the need for a more usable interface that will help people to feel more empowered.

7.2 Suggestions

In our second exit survey question, we asked: *Thinking about the simulation that you just used, what could be done to make it easier to find information about how apps access sensitive information?*

This question had two purposes: (i) to gather specific design recommendations from participants who used TurtleGuard; (ii) to get general suggestions from participants who used the default permission manager.

In total, 66.03% participants (383 of 580) suggested at least some change or improvement (Cohen's $\kappa = 0.896$). Table 4 shows the breakdown of how many participants in each condition prefer a change in the dashboard within their condition. A chi-square test shows a statistically significant association between a participant's condition and whether the participant wants changes in the dashboard ($p < 0.00005, \phi = 0.237$). This suggests the participants in the *experimental* condition are more satisfied with the controls provided by the new design than those in the *control* condition. Our work aims to fill the need users have regarding control over permissions and their personal privacy.

The most common suggestion (32.24% of all suggestions) was to reduce the number of layers to the actual permission interface (Cohen's $\kappa = 0.603$). Participants complained about number of different interfaces they had to traverse before reaching the actual permission interface. Many participants suggested that they would prefer to reach a permission control interface directly through the application—either as part of the application or by pressing the app icon. TurtleGuard addresses this concern by providing a path to permission management that involves fewer clicks and centralizes all permission management functionality.

- "*Streamline the interface to require less touches to find the information about permissions and make it explicit as to what type of data would be collected if allowed.*"
- "*Perhaps have an easier way to access the app's settings, such as holding onto an app's icon will bring up its specific settings.*"
- "*Make each app itself have the option to find that information instead of going to the general phone settings.*"
- "*There should be one centralized location, or maybe an app for that. Just to toggle with these very important settings.*"

Seven participants thought having a log of recent resource usage by applications would be useful. Some went further, mentioning that the log should also provide contextual cues, such as the visibility of the application at the time it makes the request. This finding provides evidence in support of Liu et al. [20], that recent statistics help users make better decisions. TurtleGuard provides this functionality by showing all the recent resource requests along with (i) the decision that platform took on behalf of the users, (ii) the time that the decision was made, and (iii) the visibility of the requesting application.

- "*It would be useful to have a dashboard which shows which apps are accessing what and when. Being able to see a log of the actual data that was accessed would also be useful.*"
- "*A log could be provided as an option in the settings that shows all times an app accessed sensitive information.*"

A few participants (14.6%) also suggested that there should be a tutorial, wizard style guide, or a FAQ to explain how to manage permissions (Cohen's $\kappa = 0.651$). Some wanted the applications to explain *why* they need access to certain resources. Some even suggested runtime prompts for every sensitive request access. One participant suggested that app developers hold a YouTube Q&A session on resource usage after each release:

- "*As the app is being introduced to the users, they should make a youtube q&a to answer any simple questions like this.*"

Prior work has already shown that having runtime prompts on every sensitive request is not feasible [38]—we believe that a log of recent resource accesses with surrounding context is the closest practical solution.

8. DISCUSSION

Our primary goal is to empower users to make privacy decisions better aligned with their preferences and to keep them informed about how third-party applications exercise granted permissions, and under what circumstances. We performed iterative user-centered design on a new permission management interface, TurtleGuard, which offers users significant improvements in their ability to control permissions when compared to the default permission manager.

8.1 Auditing Automated Decision Making

Recent research uses machine-learning techniques to automatically predict users' permission preferences [39, 20, 19, 21]. While machine-learning (ML) techniques have been shown to be better at predicting users' preferences [39], they are still prone to errors.

If systems are going to use ML in the future, there must be mechanisms for users to audit the decisions made on their behalves. We believe that the design we present in our study is a critical first step towards achieving that goal. Participants using TurtleGuard were better able to understand and control *when* apps have access to sensitive data than participants using the default permission manager. A substantial proportion of participants mentioned the desire to have a log that they could use to see how each application accesses sensitive resources—functionality that is missing in the default permission manager, but is provided by TurtleGuard.

8.2 Correcting Mental Models

In Task 4, we asked participants to disable access to location data when the example app, ZvW, was not actively being used, or to explain that this was not possible. We found that 72.5% of the participants in the *control* condition incorrectly believed that this was possible. Analyzing the different paths that participants in the *control* condition took while using the Android simulation, it was evident that the majority of participants did not understand the limits of the permission interface's provided functionality. This mismatch between users' mental models and actual functionality may lead to users incorrectly believing that they have denied access to certain requests for sensitive data.

8.3 Privacy Nudges

Previous work investigated ways to nudge users to configure their privacy settings and make them aware of how applications access their data [20, 13, 17]. While helping motivate users to use TurtleGuard (and other privacy management interfaces) is important, it is out of scope for this work. Nevertheless, our survey results showed that 63.1% of participants—independent of condition—previously searched for permission information on their smartphones. This shows that users are keen to understand how applications use their sensitive resources, and interfaces similar to the one we present in this study fill a critical need.

8.4 Limitations

In our proposed interface, TurtleGuard, we allow users to vary their decisions based on the visibility of the requesting application. We believe this is a significant first step towards enabling users to make contextual privacy decisions. The full extent of the impact of the surrounding context, however, goes beyond the mere visibility of the requesting application. More work is needed to understand different contextual factors and their respective impact on users' privacy decisions. We hope the results of this study will pave the path for future work on implementing *fully* contextually aware permission managers.

Additionally, we acknowledge the limitations in our screening process: participants who selected Android as their mobile device may have varying levels of usage and knowledge regarding the platform. Prior experience may have rendered the default permission manager as being easier to use for some participants in the *control* condition. This suggests that for new Android users, the usability improvements of TurtleGuard may be even greater than what we observed.

We also acknowledge that irregularities in the simulation may have had an impact towards participants' comprehension and completion rates. These confounding factors introduced by the UI, however, would have impacted both conditions equally, because the control condition was simulated using the same infrastructure and development environment. Finally, for users in the control condition, Task 4 may have been deceptively tricky due to its impossibility. Nevertheless, the incorrect answers underscore a very real problem: Android users are not aware that they are unable to deny resources to applications that they are not using.

8.5 Conclusion

Android's existing permission models, ask-on-install (AOI) and ask-on-first-use (AOFU), are insufficient at fulfilling users' privacy desires and needs. Neither of the existing models account for contextual factors in their decisions to allow or deny access to sensitive data. Users want to protect their sensitive information, but have a hard time understanding when access to data is and is not being allowed. TurtleGuard adds both ease of use and functionality, including the ability to consider application visibility when specifying privacy preferences, which has been shown to be a strong contextual cue. In our study of TurtleGuard, we had participants perform permission-related tasks and compared their performance TurtleGuard with a control group using the default permission manager. Based on our results, we iterated on TurtleGuard's design, and then performed a validation experiment to confirm the validity of our changes. Our results show that users are significantly better at performing permission management tasks with TurtleGuard than the default permission manager.

Acknowledgements

This research was supported by the United States Department of Homeland Security's Science and Technology Directorate under contract FA8750-16-C-0140, the Center for Long-Term Cybersecurity (CLTC) at UC Berkeley, the National Science Foundation under grants CNS-1318680 and CNS-1514457, Intel through the ISTC for Secure Computing, and the AFOSR under MURI award FA9550-12-1-0040. The content of this document does not necessarily reflect the position or the policy of the U.S. Government and no official endorsement should be inferred.

9. REFERENCES

[1] H. Almuhimedi, F. Schaub, N. Sadeh, I. Adjerid, A. Acquisti, J. Gluck, L. F. Cranor, and Y. Agarwal. Your location has been shared 5,398 times!: A field study on mobile app privacy nudging. In *Proc. of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 787–796. ACM, 2015.

[2] P. Andriotis and T. Tryfonas. Impact of user data privacy management controls on mobile device

investigations. In *IFIP International Conference on Digital Forensics*, pages 89–105. Springer, 2016.

[3] Apple. About privacy and location services for ios 8 and later. https://support.apple.com/en-us/HT203033. Accessed: March 4, 2017.

[4] A. Barth, A. Datta, J. C. Mitchell, and H. Nissenbaum. Privacy and contextual integrity: Framework and applications. In *Proc. of the 2006 IEEE Symposium on Security and Privacy*, SP '06, Washington, DC, USA, 2006. IEEE Computer Society.

[5] M. Benisch, P. G. Kelley, N. Sadeh, and L. F. Cranor. Capturing location-privacy preferences: Quantifying accuracy and user-burden tradeoffs. *Personal Ubiquitous Comput.*, 15(7):679–694, Oct. 2011.

[6] M. Bokhorst. Xprivacy. https://github.com/M66B/XPrivacy, 2015.

[7] CollegeDev. Donkeyguard. https://play.google.com/store/apps/details?id=eu.donkeyguard, 2014.

[8] S. Egelman, A. P. Felt, and D. Wagner. Choice architecture and smartphone privacy: There's a price for that. In *The 2012 Workshop on the Economics of Information Security (WEIS)*, 2012.

[9] A. P. Felt, E. Chin, S. Hanna, D. Song, and D. Wagner. Android permissions demystified. In *Proc. of the ACM Conf. on Comp. and Comm. Sec.*, CCS '11, pages 627–638, New York, NY, USA, 2011. ACM.

[10] A. P. Felt, S. Egelman, M. Finifter, D. Akhawe, and D. Wagner. How to ask for permission. In *Proc. of the 7th USENIX conference on Hot Topics in Security*, Berkeley, CA, USA, 2012. USENIX Association.

[11] A. P. Felt, S. Egelman, and D. Wagner. I've got 99 problems, but vibration ain't one: a survey of smartphone users' concerns. In *Proc. of the 2nd ACM workshop on Security and Privacy in Smartphones and Mobile devices*, SPSM '12, pages 33–44, New York, NY, USA, 2012. ACM.

[12] A. P. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D. Wagner. Android permissions: user attention, comprehension, and behavior. In *Proc. of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, New York, NY, USA, 2012. ACM.

[13] H. Fu, Y. Yang, N. Shingte, J. Lindqvist, and M. Gruteser. A field study of run-time location access disclosures on android smartphones. *Proc. USEC*, 14, 2014.

[14] N. Good. The Deadly Sins of Security User Interfaces. In M. Jakobsson, editor, *The Death of the Internet*, chapter 7.5, pages 398–415. John Wiley & Sons, 2012.

[15] Google. Normal and dangerous permissions. https://developer.android.com/guide/topics/permissions/requesting.html#normal-dangerous.

[16] P. Hornyack, S. Han, J. Jung, S. Schechter, and D. Wetherall. These aren't the droids you're looking for: retrofitting android to protect data from imperious applications. In *Proc. of the ACM Conf. on Comp. and Comm. Sec.*, CCS '11, pages 639–652, New York, NY, USA, 2011. ACM.

[17] L. Jedrzejczyk, B. A. Price, A. K. Bandara, and B. Nuseibeh. On the impact of real-time feedback on users' behaviour in mobile location-sharing applications. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, page 14. ACM, 2010.

[18] P. G. Kelley, S. Consolvo, L. F. Cranor, J. Jung, N. Sadeh, and D. Wetherall. A conundrum of permissions: Installing applications on an android smartphone. In *Proc. of the 16th Intl. Conf. on Financial Cryptography and Data Sec.*, FC'12, pages 68–79, Berlin, Heidelberg, 2012. Springer-Verlag.

[19] H. Lee and A. Kobsa. Privacy Preference Modeling and Prediction in a Simulated Campuswide IoT Environment. In *IEEE International Conference on Pervasive Computing and Communications*, 2017.

[20] B. Liu, M. S. Andersen, F. Schaub, H. Almuhimedi, S. A. Zhang, N. Sadeh, Y. Agarwal, and A. Acquisti. Follow my recommendations: A personalized assistant for mobile app permissions. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, 2016.

[21] B. Liu, J. Lin, and N. Sadeh. Reconciling mobile app privacy and usability on smartphones: Could user privacy profiles help? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 201–212, New York, NY, USA, 2014. ACM.

[22] W. Mason and S. Suri. Conducting behavioral research on amazon's mechanical turk. *Behavior Research Methods*, 44(1):1–23, 2012.

[23] D. Mate. Permission master. https://play.google.com/store/apps/details?id=com.droidmate.permaster, 2014.

[24] M. McLaughlin. What is lineageos. https://www.lifewire.com/what-is-cyanogenmod-121679, 2017.

[25] H. Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79:119, February 2004.

[26] K. Opsahl. Uber should restore user control to location privacy. https://www.eff.org/deeplinks/2016/12/uber-should-restore-user-control-location-privacy, 12 2016.

[27] G. Paolacci and J. Chandler. Inside the turk. *Current Directions in Psychological Science*, 23(3):184–188, 2014.

[28] S. Patil, Y. Le Gall, A. J. Lee, and A. Kapadia. My privacy policy: Exploring end-user specification of free-form location access rules. In *Proceedings of the 16th International Conference on Financial Cryptography and Data Security*, FC'12, pages 86–97, Berlin, Heidelberg, 2012. Springer-Verlag.

[29] S. Patil, R. Schlegel, A. Kapadia, and A. J. Lee. Reflection or action?: How feedback and control affect location sharing decisions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 101–110, New York, NY, USA, 2014. ACM.

[30] E. Peer, L. Brandimarte, S. Samat, and A. Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163, May 2016.

[31] D. Reilly, D. Dearman, V. Ha, I. Smith, and K. Inkpen. "need to know": Examining information need in location discourse. In *Proceedings of the 4th International Conference on Pervasive Computing*, PERVASIVE'06, pages 33–49, Berlin, Heidelberg, 2006. Springer-Verlag.

[32] X. M. Repository. http://repo.xposed.info/, http://repo.xposed.info/.

[33] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomlinson. Who are the crowdworkers?: Shifting demographics in mechanical turk. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, pages 2863–2872, New York, NY, USA, 2010. ACM.
[34] J. L. B. L. N. Sadeh and J. I. Hong. Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings. In *Symposium on Usable Privacy and Security (SOUPS)*, 2014.
[35] F. Shih, I. Liccardi, and D. Weitzner. Privacy tipping points in smartphones privacy preferences. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 807–816, New York, NY, USA, 2015. ACM.
[36] I. Shklovski, S. D. Mainwaring, H. H. Skúladóttir, and H. Borgthorsson. Leakiness and creepiness in app space: Perceptions of privacy and mobile app use. In *Proc. of the 32nd Ann. ACM Conf. on Human Factors in Computing Systems*, CHI '14, pages 2347–2356, New York, NY, USA, 2014. ACM.
[37] C. Thompson, M. Johnson, S. Egelman, D. Wagner, and J. King. When it's better to ask forgiveness than get permission: Designing usable audit mechanisms for mobile permissions. In *Proc. of the 2013 Symposium on Usable Privacy and Security (SOUPS)*, 2013.
[38] P. Wijesekera, A. Baokar, A. Hosseini, S. Egelman, D. Wagner, and K. Beznosov. Android permissions remystified: A field study on contextual integrity. In *Proceedings of the 24th USENIX Conference on Security Symposium*, SEC'15, pages 499–514, Berkeley, CA, USA, 2015. USENIX Association.
[39] P. Wijesekera, A. Baokar, L. Tsai, J. Reardon, S. Egelman, D. Wagner, and K. Beznosov. The feasibility of dynamically granted permissions: Aligning mobile privacy with user preferences. *arXiv preprint 1703.02090*, 2017.

APPENDIX

Permission	Explanation
CALL_PHONE PROCESS_OUTGOING_CALLS READ_PHONE READ_CALL_LOG ADD_VOICEMAIL WRITE_CALL_LOG	Make and process calls as well as read information about call status, network information and previously made phone calls
CAMERA	Access camera devices
GET_ACCOUNTS	Access to list of accounts
READ_CALENDAR WRITE_CALENDAR	Read and write events to the user's calendar
READ_CONTACTS WRITE_CONTACTS	Read and write to user's contacts
READ_EXTERNAL_STORAGE WRITE_EXTERNAL_STORAGE	Read and write files to the user's external storage
RECORD_AUDIO	Record audio
ACCESS_COARSE_LOCATION ACCESS_FINE_LOCATION ACCESS_WIFI_STATE	Read location information in various ways including network SSID-based location
READ_SMS SEND_SMS RECEIVE_SMS	Read SMS messages from the device (including drafts) as well as send and receive new ones SMS

Table 5: Sensitive permissions managed by TurtleGuard. Permissions grouped by a single explanation form the families used in our system to reduce the number of managed permission as discussed in Section 3.

Condition	Correct	Incorrect	All
Task 1			
control	2	3	2
experimental	2	4	2
Task 2			
control	2	3	3
experimental	2	3	2
Task 3			
control	2	4	3
experimental	2	3	2
Task 4			
control	4	2	3
experimental	2	2	2

Table 6: Median ease-of-use Likert scores for all tasks, conditions, and correctness in the validation experiment. Higher scores indicate more difficulty.

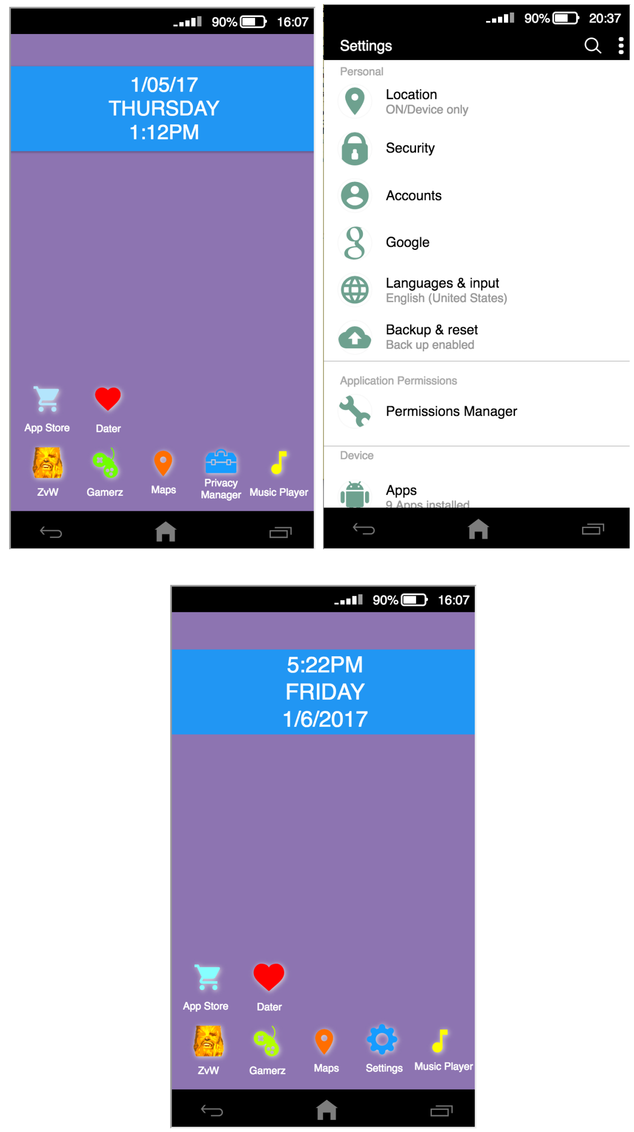


Figure 6: In the pilot experiment, TurtleGuard was launched via the icon labeled "Privacy Manager" (top left), but then added as a sub-panel to the Settings app, labeled "Permissions Manager," for the validation experiment (top right). In the *control* condition in the pilot experiment and both conditions in the validation experiment, the Settings app was accessible from the home screen (bottom).

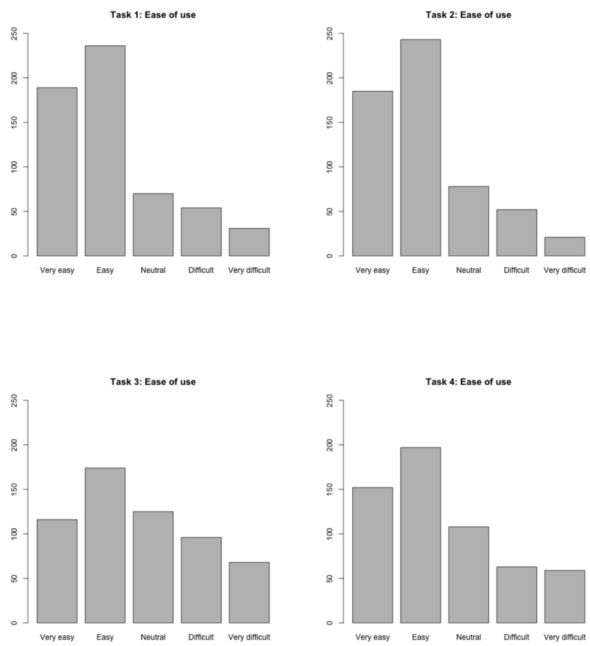


Figure 7: Ease of use histograms for each task (validation experiment)

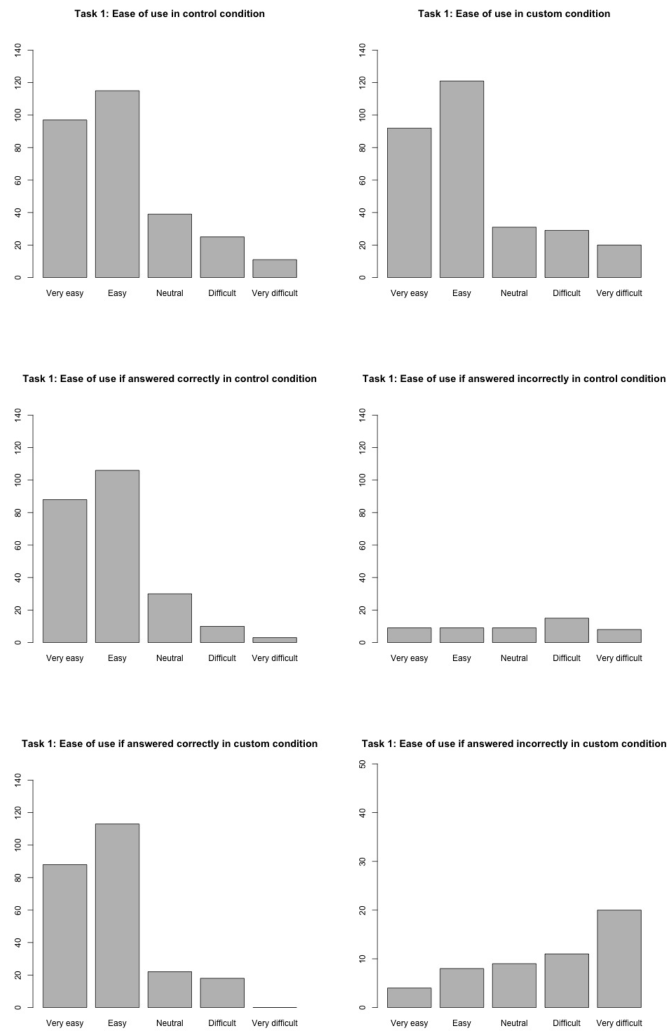


Figure 8: Ease of use histogram for Task 1 (validation experiment)

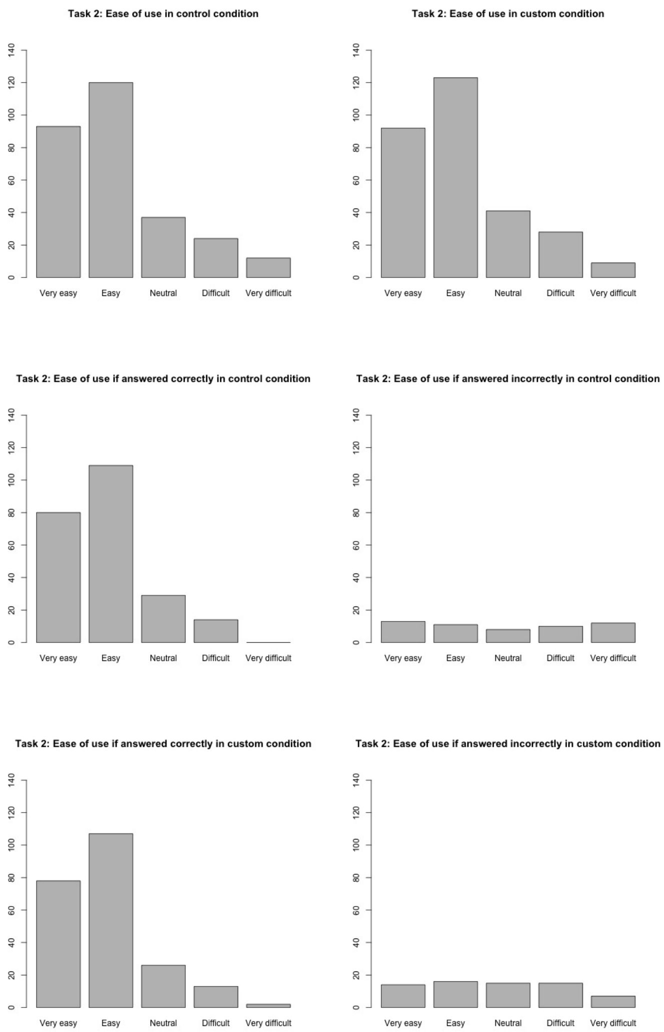


Figure 9: Ease of use histogram for Task 2 (validation experiment)

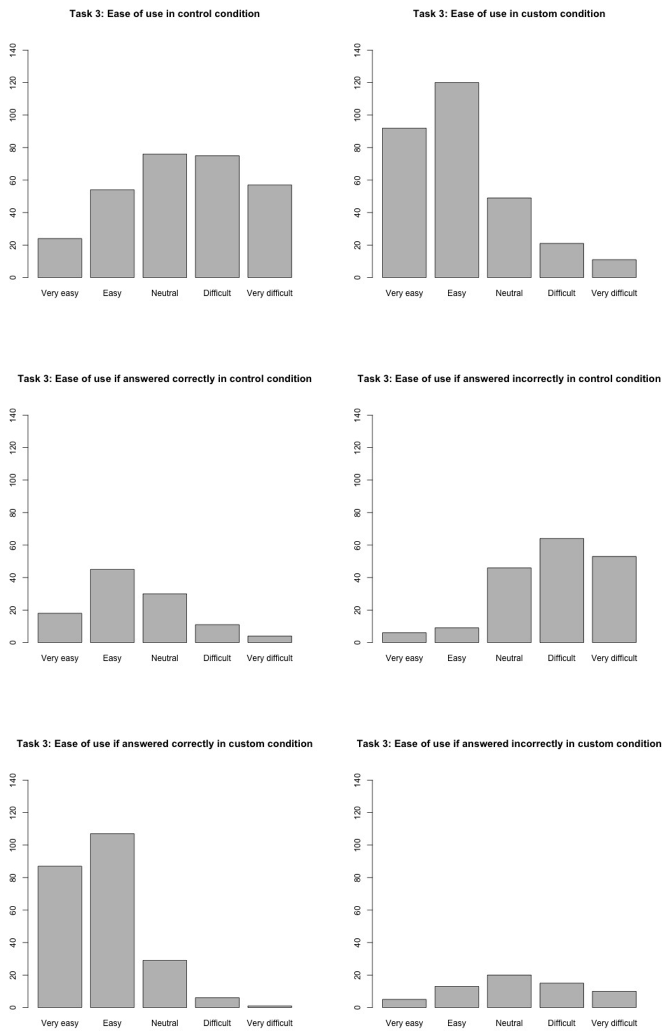


Figure 10: Ease of use histogram for Task 3 (validation experiment)

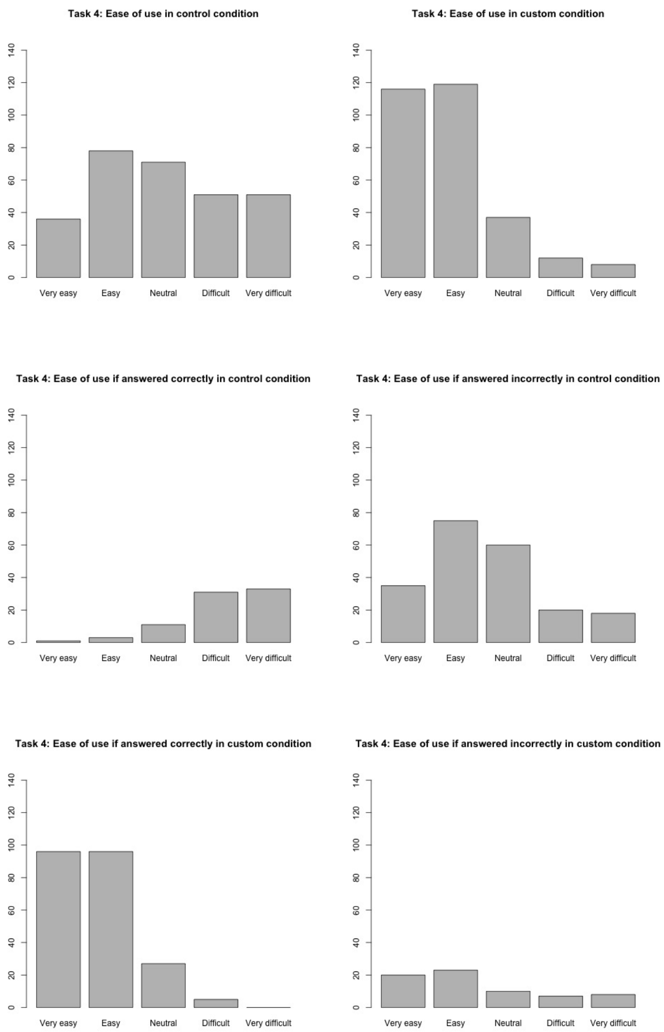


Figure 11: Ease of use histogram for Task 4 (validation experiment)

18

Authentication on the Go: Assessing the Effect of Movement on Mobile Device Keystroke Dynamics

Heather Crawford
Harris Institute for Assured Information
Florida Institute of Technology
Melbourne, FL
hcrawford@fit.edu

Ebad Ahmadzadeh
School of Computing
Florida Institute of Technology
Melbourne, FL
mahmadzadehe2012@my.fit.edu

ABSTRACT

Transparent authentication based on behavioral biometrics has the potential to improve the usability of mobile authentication due to the lack of a possibly intrusive user interface. Keystroke dynamics, or typing behavior, is a potentially rich source of biometric information for those that type frequently, and thus has been studied widely as an authenticator on touch-based mobile devices. However, the typing-while-moving scenario that characterizes mobile device use may change keystroke-based patterns sufficiently that typing biometrics-based authentication may not be viable. This paper presents a user study on the effects of user movement while typing on the effectiveness of keystroke dynamics as an authenticator. Using the dynamic text-based keystroke timings of 36 study participants, we first show that naïvely measuring patterns without considering position (e.g., sitting, standing or walking while typing) results in generic patterns that are little better than chance. We show that first determining the user's position before classifying their typing behavior, our two-phased approach, inferred the user's position with an AUC of above 90%, and the user's typing pattern was classified with an AUC of over 93%. Our results show that user typing patterns are a viable secondary or continuous post-PIN authentication method, even when movement changes a user's typing pattern.

1. INTRODUCTION

Mobile devices have become full computing platforms. The data and services they provide have made protecting them of paramount importance. Most devices use a secret knowledge-based means to protect them, such as a password, PIN, or small sketch (e.g., Android pattern lock). These are appropriate measures for initially protecting the device, but they do not provide protection if the device owner does not use them, or if an attacker gains access to an unlocked device. Keystroke dynamics, or the way in which a person types, has been suggested as a possible means to improve authentication by allowing it to be both *continuous*, protecting the device even after the initial password has been entered, but also *transparent* in that the user need not be distracted from their main task in order to authenticate regularly [42]. This has the potential to not only provide a higher device security level by continuing to authenticate the user after initial password entry, but also improve usability by removing a potentially disruptive request for repeated authentication.

Many of the existing keystroke dynamics studies have relied on the user typing a fixed word or phrase, such as adding keystroke dynamics to password entry, a practice known as *password hardening* [35], but not on dynamic text that changes from sample to sample. Also, much effort has gone into selecting the "best" classifier or the "best" set of features, with only small changes in the apparent distinctive nature of either.

This paper presents a keystroke dynamics user study designed to determine whether user typing patterns change enough during movement that it can no longer be used as an authenticator. We found on initial analysis that typing patterns over three positions (sitting, standing and walking) were insufficiently distinct to be used as evidence for authentication. This poor result is due to the additional movement that classification algorithms must overcome while typing. We have developed a phased classification approach, seen in Figure 1, that takes advantage of such movement. Our phased approach begins with using gyroscope data gathered at each keypress to determine the user's position (sit, stand, or walk). Next, classification models are created for each of the three positions under study that are then used to classify new data. The work presented here is a feasibility study to determine whether the collected gyroscope data is suitable for determining user position. The main novelty in our work is showing that modeling user typing based on their position improves classification rates over building a single, position-independent model. Our results show an improvement in AUC from 66% to 97% when position is considered before classifying keystroke dynamics data. These results indicate that our phased approach has merit; future work includes simulating the classification model to determine its use in practice.

2. BACKGROUND AND RELATED WORK

2.1 Mobile Device Authentication

In addition to existing password- and PIN-based authentication methods, research has begun to emerge on alternative authentication methods that consider the mobile device's needs more closely; in particular interest in using graphical passwords as an authenticator has been demonstrated [38, 41]. However, these methods still provide an all-or-nothing

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

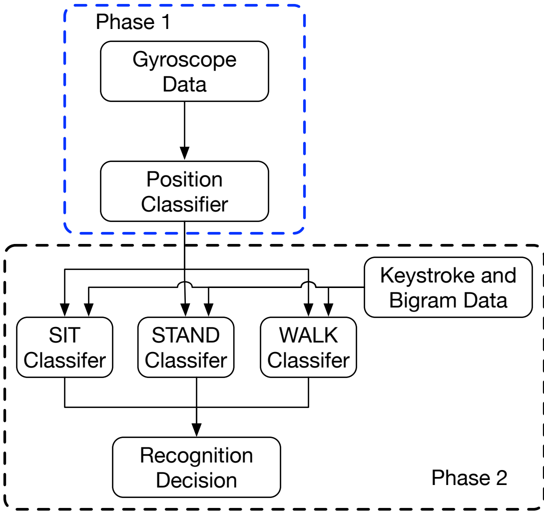


Figure 1: Our two-phased classification model

approach to device protection in that once the user is correctly authenticated, they are granted access to all data, services and apps on the device. In response, researchers have begun to study methods that continue to authenticate the user invisibly in the background while other tasks are completed. This is called *transparent, continuous* authentication. This type of authentication gathers behavioral biometric data such as keystroke dynamics [9, 32], touches [20, 42], etc. to continuously ensure that the device owner is the one currently using the device. Methods by which access to apps, data and services on the device can be restricted based on the identity of the current user have also begun to emerge [19].

2.2 Keystroke Dynamics

Keystroke dynamics is a behavioral biometric that uses patterns in how a person types to distinguish them from other users. It uses metrics such as *key hold time* (the amount of time between pressing and releasing the same key) and *inter-key latency* (the amount of time that passes between releasing one key and pressing the next) to identify these distinctive patterns. Researchers have examined other potential biometrics such as touch [11], facial recognition [15] and device movement [17]. Keystroke dynamics began with studies on desktop and laptop computers [30, 36] and in recent years has moved to mobile devices such as smartphones [13, 32]. Many keystroke dynamics studies attempt to replicate a password hardening situation in which the data gathered during the study is based on a known password that each study participant types a set number of times [28]. This practice can increase the strength of traditional passwords, but still provides an all-or-nothing approach to authentication. More recently, research has focused on providing transparent authentication that protects throughout device use rather than just at the beginning. It is this research that maps most directly to ours; thus we will focus on the current work in this area.

Typing patterns while moving have been studied by Clawson *et al.* in an effort to determine whether moving while typing affects accuracy and errors [16]. Their study had 36 participants type set phrases using a hard keyboard on a mobile device (Blackberry Curve 8320) while walking a set path. Their results showed that expert typists made fewer mistakes while walking, but at the cost of a lower typing speed [16]. Clawson *et al.*'s work is supportive of ours since more accurate typing may lead to improved uniqueness in typing patterns. However, Clawson *et al.* were not studying the use of keystroke dynamics as an authentication method, so further comparison of our results to this work are not indicated.

There has been much discussion on the amount and type of text used as input to transparent keystroke dynamics authentication tools. Many of the studies in this area, specifically those to do with password hardening, focus on text that must be repeated, while continuous, transparent authentication methods are likely to be based on any text the user may type. There is also the need for ecological validity – if a user can be expected to type any words and phrases, then basing a user study on specific words or phrases cannot be used to justify results in a more open environment.

2.2.1 Fixed Text

Fixed text methods (also called *static text*) assume that the user will type the same word or phrase at both enrollment and at the time of authentication. The text typed is generally short, as typing long texts at the time of authentication is tedious and error-prone on a mobile device [2, 12]. In general, using fixed text allows for more stability as the comparison between enrolled sample and gathered sample share the same keys and are thus similar. In some cases, experiments of this type produce results that either depend on special conditions (such as the attacker knowing the user's password) or have unacceptable accuracy levels [23, 7]. Much research has been done on fixed text methods [14, 28]; a summary of work in this area may be found in [18].

2.2.2 Dynamic Text

Also called *free text*, this paradigm assumes that the user may type whatever they wish, and that this input is any length. In reality, dynamic text and free text have several differences; free text is completely without constraints, where dynamic text may have aspects of both fixed and free text. Specifically, dynamic text may be prompted in some way, or may depend on a small number of specific words or phrases [23]. Several studies have examined dynamic text keystroke dynamics [4, 34], including Ahmed *et al.* [1] who report fairly good results, with False Accept Rates well below 1%. Free text keystroke dynamics has also been studied by Gunetti & Picardi [23], although their reported results are not as low as those of Ahmed *et al.* A summary of work on free text keystroke dynamics is available in [3]. The implication is that transparent authentication based on keystroke dynamics is best suited to *true* free text, which removes any restrictions about what or how much is typed. In this way, any characters a user may type can potentially be used as information upon which to base authentication decisions.

Adding realism to mobile device keystroke dynamics experiments has been studied from several points of view. One is that users may change their hand positioning while typing, which may affect their overall typing pattern. Azenkot & Zhai [6] studied user typing patterns when typing with one thumb, both thumbs and one index finger and found that there were pattern differences between these three hand positions. They used these results to suggest changes in keyboard design and layout that can improve typing accuracy. Similarly, Buschek *et al.* [10] studied the same three hand positions, but from the point of view of authentication rather than keyboard improvements. Their results showed that hand position had a strong effect on the ability to authenticate a user [10]. Both of these papers were based on password-hardening techniques, and thus were using fixed text techniques with defined feature vectors. Shen *et al.* studied the use of motion sensor data while typing a mobile device passcode as a potential authenticator [39]. They reported a False Reject Rate (FRR) of 6.85% and a False Accept Rate (FAR) of 5.01% in a user study with 48 participants [39]. Their work is similar to ours since they report results in the seated, standing and walking positions, although they only consider the sensor data when unlocking the mobile device with a passcode.

2.3 Gyroscope Data

Modern mobile devices come equipped with built-in sensors that can measure motion, orientation, environmental conditions such as temperature and humidity, etc. Data from sensors such as accelerometers and gyroscopes has been used for activity recognition [8, 25], to address typing inaccuracies [22], to create keyloggers [33], and to determine on-device input errors [37]. Accordingly, authentication research has begun to consider whether accelerometer and gyroscope data may be used as a unique identifier. Giuffrida *et al.* created what they call "sensor-enhanced" keystroke dynamics in their UNAGI system [21]. They experimented with the use of accelerometer and gyroscope data while 20 participants typed a set of fixed passwords and found that they were able to achieve Equal Error Rate (EER) values of less than 1% [21]. Their use of a fixed password as the stimulus indicates that theirs was a password hardening experiment rather than a dynamic text experiment.

2.4 Contribution

The major contribution of this paper is the determination that while typing patterns do change with user movement while typing, our phased classification model allows keystroke dynamics to be used as a viable secondary authentication method under realistic movement and text-acquisition conditions despite typing changes. To our knowledge, this is the first work to create different keystroke models for different user positions; a step that improves the accuracy of user classification. We also provide evidence that gyroscope data gathered at the time of each keypress is suitably distinctive to distinguish between sitting, standing and walking positions. This is significant because gyroscope data is often sampled essentially continuously, which generates a lot of data, uses significant battery power, and requires significant processing in order to be useful. Overall, our results show that keystroke dynamics can be used as secondary or continuous authentication method.

3. RESEARCH QUESTIONS

Our research questions are as follows:

1. Does gyroscope data captured at the time keypresses were made provide enough information to tell whether the typist is seated, standing or walking?
2. Does creating multiple position-specific models for a typist provide better classification results compared to using a single model trained on all positions?
3. Does a dynamic text-based system based on the above assumptions provide enough data for a sufficiently distinctive user model?

3.1 Hypotheses

Based on the above research questions, we present the following hypotheses for our work:

Hypothesis 1: A mobile device user's typing pattern is distinctive enough to use as a secondary or continuous authentication method as determined by achieving an AUC of at least 90%.

Hypothesis 2: Gyroscope data gathered as a key is pressed is distinctive enough to determine whether the typist is seated, standing or walking while typing, as determined by achieving an AUC of at least 90%.

Hypothesis 3: Determining a user's position and classifying based on data from that position only decreases False Accept and False Reject Rates (FAR and FRR, respectively) when compared to classification without determining user position.

We chose 90% as the AUC for Hypothesis 1 in order to justify using our method as a secondary or continuous authentication method, e.g., one that takes place as a supplement to or after primary authentication such as via a password or PIN. This means that near-perfect accuracy is not required, and the balance between FAR and FRR is not as vital as for a primary authentication method. While we may have chosen a lower AUC, we wish to produce a system that may be viable for primary authentication in the future. Therefore, 90% AUC is a value balanced between these two design choices. We chose AUC of 90% for Hypothesis 2 because high accuracy is not required since position determination is not an authentication decision (although it is related to one via our two-phased approach) and thus does not have the security ramifications that characterize authentication decisions.

4. THREAT MODEL

We assume that an attacker has access to the unlocked mobile device and may have had an opportunity to observe the device owner typing, and thus would know things such as current position, preferred hand position (e.g., index finger, one thumb, both thumbs), device orientation (e.g., landscape or portrait) and a general idea of typing speed. The attacker is assumed to have full knowledge of the biometric authentication system, including all inputs and outputs.

5. STUDY AND DATA COLLECTION

We collected gyroscope data and dynamic text typing data in a user study in a single session. The participant used a custom-built Android app to type phrases provided to them that varied both their position and the device orientation while typing. Specifically, participants were prompted by the experimenter to hold the device in a given orientation (portrait or landscape) and to type while either seated, standing or walking. The participants were told to type as they usually did; specifically, the speed of their typing was not restricted. We also did not provide specific guidance on how to sit, stand or walk. For instance, many participants chose to stand while leaning against a wall, or sit with their arms supported by a table. The only prompts we gave during the experiment were to keep walking if the participant stopped while typing in a walking condition. The study participants filled out a short demographic questionnaire before beginning any typing, and they were allowed to rest between conditions if they wished. Each participant was given the opportunity to practice typing before beginning the first condition; this training data was discarded before analysis. This study was approved by our university's Institutional Review Board (IRB) prior to its start.

5.1 Participants

We recruited 39 participants (6 female, 33 male) through convenience sampling methods such as personal invitation, emails to mailing lists within our university and word of mouth. The data from three participants was removed from the study due to procedural errors, leaving data from 36 participants (5 female, 31 male). The remainder of this paper, including study results, refers to the analysis of data from the remaining 36 participants. The average age of participants was 28.3 years (SD = 11.3 years). Participants were not required to have any experience with typing on smartphones, although all participants reported that they owned and used a mobile device, most with soft keyboards. 2 participants were left-handed, and 34 were right-handed. Participant experience on their own mobile device varied: 14 participants used an Android-based device, 18 used an iOS device, 2 used another smartphone, and 2 used a feature (non-smart) phone. 2 participants were considered novices (used their device once a week or less), 3 participants as average (used their device more than once a week but not everyday) and 31 participants as experts (used their device every day or several times each day). Most participants were students, faculty or staff at our university; all participants had at least some post-secondary education, ranging from some undergraduate experience to graduate levels. Participants were not compensated for their participation.

5.2 Apparatus

We provided each participant with an LG Nexus 5 smartphone for the duration of the experiment. Each device ran Android version 4.4.4 and contained only the standard Android applications. Text entry was facilitated by the use of two bespoke Android applications. The first (see Figure 2) displayed the phrase to be typed (non-editable), a text box where the user typed the same phrase, and a counter that displayed the number of phrases the participant had typed in the current experimental condition. This app randomly chose a phrase from a modified version of the standard phrase set provided by MacKenzie and Soukoreff [31] (forthwith called the M&S set); duplicate phrases were permitted.

The second Android app used in this study (see Figure 3) was a custom-designed keyboard. It was designed to visually mimic the standard Android keyboard in order to accurately emulate a standard typing environment; the same keyboard design was used by all participants. This app was responsible for gathering the required keystroke and bigram metrics. When the participant pressed a key, the app recorded the key pressed, key hold time, inter-key latency, device orientation, user position and instantaneous gyroscope data (pitch, azimuth and roll). Key hold time is defined as the amount of time that a participant holds down a given key. Inter-key latency is defined as the amount of time between a key up event and the subsequent key down event.

Figure 2: Phrase generation app screenshot

Figure 3: Custom keyboard for metric gathering

Timing of such typing events is a subject of debate in the keystroke dynamics field [27] as incorrect timing accesses can affect the measured typing pattern of a participant, which in turn has an effect on the reported study results. We mitigated such potential sources of error by using a set of four devices of the same model with the same operating system build, all of which had been reset to factory settings before the experiment began. In addition, we used the same Android applications on each device, and removed the previous participant's gathered data and restarted the application between participants. By using these precautions, we have made all possible efforts to minimize the effects of clock discrepancies on the results of this study.

Our keyboard, which runs as a service on the Nexus 5 devices, replaced the default keyboard in the settings of each device. This design means that when the user tapped on a widget that can accept keystrokes, our keyboard was displayed and subsequently used by the study participants. This allowed for gathering keystrokes from all applications that required typing, meaning that this same keyboard could be used in future work on transparent keystroke dynamics-based authentication.

5.3 Study Procedure

Each participant first answered a short demographic questionnaire, then was introduced to the app and bespoke soft keyboard they would use for phrase entry. They were given a choice to practice with the standard Android keyboard if they were unfamiliar with it. Most declined as they felt they had enough experience typing on the standard soft keyboard. The participants were allowed to take short breaks after each experimental condition. The participants were instructed to type in their usual manner, and that speed or accuracy were not being measured. They were told that auto-correction and auto-capitalization were disabled, and that if they made mistakes it was their decision whether or not to correct them. The participants were told to not change the device orientation and to remain in the participant mode (sit, stand, walk) they were placed into by the experimenter. They were not told how fast to walk, nor whether they should (or should not) support the device while typing (i.e., leaning against a wall, or with arms supported on a tabletop while seated). No further specifications were given to participants.

Each participant was placed into each of six experimental conditions (see Section 5.5 for a of the conditions) by the experimenter and asked to return to the experimenter when they had typed at least 22 phrases (there was a counter at the top of the custom phrase app for this purpose). The number of phrases was chosen in order to gather enough data for analysis, and to provide a similar amount of data for each participant. After completing each condition, the participant was asked to return to the experimenter, who would place the device into the next condition and instruct the participant as to their mode and the device orientation (i.e., "Please type the next set of phrases with the device in portrait while you're seated"). Once the participant had completed each of the six conditions, they were thanked for their time and allowed to leave.

5.4 Phrase Sets

The stimulus item in this experiment – the prompt that encouraged the user to type – was a randomly selected phrase from a modified set of standard phrases (the M&S set) [31]. Much debate has ensued over the choice of phrases used in text entry experiments. The main issues are that having a non-standard phrase set may impact the results of the study in that using a different phrase sets may result in different experimental results [29]. MacKenzie & Soukoreff addressed the issues of experimental validity (both internal and external) [31], and provided a set of 500 phrases that have been used in various studies. Kristensson & Vertanen opine that the phrase set chosen has an effect on study reproducibility in addition to internal and external validity; it is nearly impossible to reproduce an experiment if the actual phrase set is unknown (i.e., taken from random selections from an unspecified source, such as collecting phrases from "the news") [29].

In choosing a phrase set for this experiment, we kept in mind both internal and external validity as well as study reproducibility, while ensuring our phrase set met the requirements of our experiment. Specifically, we required a phrase set that closely matched letter frequencies in English, was large enough to ensure repeated phrases for the same participant were minimized, and contained upper- and lower-case letters, punctuation and numbers. The M&S set met the first two requirements; but not the last one. To remedy this, we edited the M&S set to have upper-case letters at the start of each phrase, changed text numbers to numeric equivalents (i.e., "eight" was changed to "8"), and added punctuation such as ending periods, exclamation points and question marks, as well as commas where grammatically correct. We believe that doing so created a phrase set that was both ecologically valid and made for a repeatable experiment.

Typically, a true free text typing experiment would require the participant to type whatever came to mind. One issue that arises, though, is what to do when the user cannot think of anything to type since this will affect their standard typing pattern. The role of the phrases in our study were to keep the user typing in as natural a way as possible. Otherwise, the content of the phrases did not have an impact on the accuracy, difficulty, or usability of the typing task. By using phrases rather than free text, however, we have change the user's task from one of creation to one of transcription, which may have an impact on their typing pattern. The advantage we gain is that the typing data is captured with a higher degree of freedom and fewer restrictions than with comparable fixed text experiments, which is arguably more similar to real-world typing situations.

5.5 Experiment Design and Analysis

Our laboratory-based study used a within-subjects, repeated measures design, in which the study participants were assigned to one of six experimental groups that differed only in the order in which the participant completed each of the six study conditions (see Table 1). All participants completed all study conditions. Participants were assigned to each study condition using a $6x6$ latin squares design in order to minimize learning effects and fatigue. Each session lasted about one hour.

<table>
<tr><th colspan="3">Study Conditions</th></tr>
<tr><th>Position</th><th>Device Orientation</th><th>Description</th></tr>
<tr><td>Sit</td><td>Portrait</td><td rowspan="2">Sitting in a fixed chair (no casters);
Arms optionally supported on table;</td></tr>
<tr><td>Sit</td><td>Landscape</td></tr>
<tr><td>Stand</td><td>Portrait</td><td rowspan="2">Standing, device unsupported;
Optionally leaning against a wall;</td></tr>
<tr><td>Stand</td><td>Landscape</td></tr>
<tr><td>Walk</td><td>Portrait</td><td rowspan="2">Walking around a large space, some obstacles;
No set speed; most users walked slowly</td></tr>
<tr><td>Walk</td><td>Landscape</td></tr>
</table>

Table 1: Description of study conditions

5.6 Data Gathered

The collected keystroke data was sanitized by removing the %, & and $ characters because the experimenter long pressed these keys to indicate a transition between the six study conditions. We used these keys as indicators of a change in device mode between sit, stand and walk. We chose these three keys for this purpose because they did not appear in any of the 500 phrases used as stimulus items, and thus could safely be removed from the dataset without removing valuable user data.

For bigram data, we collected the two characters that make up each bigram (not used during data analysis), the calculated key hold time for that bigram, the device orientation (portrait or landscape) and the participant's position (sit, stand, or walk). We sanitized the data to once again remove the occurrences of the %, & and $ characters. In the case of bigrams, we removed the entire bigram from the dataset if any of these three characters appeared as either of the two letters saved. For both keystrokes and bigrams, values greater than 3 SD beyond the mean were considered outliers and removed from the dataset prior to classification.

The gyroscope data was sanitized to remove the occurrences of %, & and $ but was unchanged otherwise. Since we gathered the gyroscope azimuth, pitch and roll in an instantaneous (i.e., at the moment of a keypress) rather than continuous manner, it was not necessary to window the data into discrete sections, nor to filter the data to remove high- or low-level frequencies as is common in activity recognition studies. Furthermore, since our gyroscope data is not time-scale data since it is discrete rather than continuous measurements, it was not necessary to transform it to the frequency domain before analysis.

We collected a total of 323,064 keystrokes and 289,520 bigrams from all 36 participants, not including practice phrases. The average number of keystrokes gathered per user was 8,974, and the average number of bigrams was 8,042. Since we gathered gyroscope data on each keystroke, we gathered the same amount of gyroscope data as keystrokes with one exception: we did not record instances of using backspace in the keystroke data, but we retained this information for gyroscope data since we were interested in the device movement on each keypress rather than whether that movement was related to a particular key.

5.7 Feature Vectors

The feature vector for the gyroscope data was simply the x, y, and z coordinates as gathered during each keypress. The makeup of a feature vector for keystrokes and bigrams, however, is much more complex.

In fixed text keystroke dynamics studies, the feature vector used is quite clear – it is the concatenation of the n key hold times corresponding to the keys pressed when typing the password (often with the ending enter keystroke) and the $n-1$ inter-key latencies for the associated bigrams. Since all participants type the same password during a study, the feature vectors are the same for each pattern gathered from each participant; the data is complete without missing values. When used outside an experimental setting, the only comparison is between a person's enrolled keystroke metrics when typing their password, and the subsequent keystroke metrics when typing the same password at a later date.

Keystroke biometrics based on dynamic text are more useful when the goal is to gather keystroke information unobtrusively, such as when continuous, transparent authentication is used to verify the identity of a person after initial login. In this situation, we specifically do not want to interrupt the user in what they are doing in order to retype their password, so we instead gather their keystroke metrics as they type as part of their regular device use. We may gather data from them when typing an email, a paper, or a blog post, all of which will have few phrases that appear in all. We gather this data from a custom extension of Android's standard keyboard so that key hold time and inter-key latency, which depend on the keyboard size and key placement, are collected in the same manner for all study participants.

Since dynamic keystroke biometrics cannot depend on getting a fixed amount of text from each participant, nor guarantee that all participants will type the same values, deciding upon the components of the feature vector is a complex task. Intuitively, selecting the most frequently typed characters and bigrams suggests that the most data possible will be retrieved from each participant. However, in practice the most frequently typed characters may vary from person to person. If the most frequently typed English letters are chosen, there might be gaps in our gathered patterns if the participant did not type that letter. This situation gets far worse when considering the frequency of bigrams. These gaps create a much more sparse dataset upon which to base authentication decisions, so far more data must be gathered to be as predictive and suitable as fixed text keystroke dynamics. To counter this, we used a dynamic feature space where the bigrams and keystrokes involved are those for which we have at least a few instances from the user. For example, early in the data collection process, the classifier may start with a minimum number of bigrams and keystrokes that have been typed thus far (we set this at 4 of each). The feature space then grows as more data is collected. For short text in which all features do not appear, we stochastically estimate the missing values from the user's past typing data.

6. RESULTS AND DISCUSSION

We now present our study results and related them back to the hypotheses defined in Section 3.1.

6.1 Position Independent Results

We begin by reporting the results of the naïve method, in which we do not use the gyroscope data to first determine user position. In this case, we mixed data from the sit, stand and walk positions and classified only based on the key hold times and inter-key latencies for two classification algorithms: Decision Tree and Logistic Regression. We chose Decision Tree because of its use in human activity recognition studies [5, 24] and because it is quick to train and classify data. Logistic regression was chosen for its simplicity and ease in understanding feature significance and removing those found to be insignificant. Furthermore, like Decision Tree, logistic regression has a low computation load for training and classifying data, which is an important feature on the constrained memory, power and processing environment on mobile devices. We considered each participant in turn the device owner (their data was considered the positive class), and the other participants as non-owner (their data was considered the negative class). The owner's model was trained on 2/3 of their supplied key hold times (keystrokes) or inter-key latencies (bigrams) plus an equal amount of data randomly selected from the other study participant's data. We used 10-fold cross-validation and report the averages from the 10 folds in Table 2. We have reported False Accept Rate (FAR), False Reject Rate (FRR), and the Area Under the Curve (AUC) for the Receiver Operating Characteristic (ROC) curve. We chose to report AUC because it provides, in a single value, the ability of our classifier to distinguish between owner typing patterns and those of others. An AUC value equal to 50% represents a method that is no better than chance; an AUC value equal to 100% is indicative of a perfect classifier.

As can be seen in Table 2, the FAR and FRR are very high for both keystrokes and bigrams. For instance, the FAR value of 41.9% for keystroke results using DT indicates that there is a 41.9% probability that an attacker will gain access. This is unacceptably high for any authentication system since it means that nearly half of all attackers will gain access to the mobile device. Similarly, the FRR of 23% for keystrokes using DT represents a nearly one in four likelihood that a legitimate user will be forced to reauthenticate. While reauthentication is less risky in terms of security, it represents an annoyance to users and a reduction in system usability since a legitimate user will have to reauthenticate once out of every four attempts.

The AUC values in Table 2 are not much better. Values in the 60-69% range represent a classifier that is only 10-19% better than chance, which is not acceptable even for secondary authentication. Overall, these results indicate that a person's typing pattern changes sufficiently over the three studied positions (sit, stand, walk) that much of the uniqueness in those typing patterns is lost.

Due to these uninspiring results, we chose not to combine key hold time and inter-key latency features as a way of improving classification rates in favor of a potentially better solution: our dual-phased classification model, which is based on first determining the user's position, then classifying using a model built using only user data from that position.

	Metric	Classifier Metric (%)		
DT		**FAR**	**FRR**	**AUC**
	Keystrokes	41.9	23.0	66.9
	Bigrams	49.3	30.5	60.3
LR		**FAR**	**FRR**	**AUC**
	Keystrokes	39.0	35.6	66.2
	Bigrams	43.3	41.7	60.7

Table 2: FAR, FRR and AUC (%) averaged over all participants for keystroke data (key hold time) and bigram data (inter-key latency) using Decision Tree (DT) and Logistic Regression (LR) classifiers. These results do not consider user position (e.g., sit, stand or walk) and are used as a baseline for comparison purposes.

6.2 Position Dependent Results

The first phase of our two-phased approach is to determine the user's position while they are typing, then classify their typing into owner or not owner based on a model trained only on data from that position. To determine position, we gathered gyroscope data from the mobile device at the moment each key was pressed. Our intuition is that the gyroscopic movement (as measured by the device's pitch, azimuth and roll) will be different when typing while seated, standing or walking. We chose not to measure accelerometer data since it is likely that the accelerometer readings will be different for the walking condition and relatively similar for seated and standing, thus making the latter two positions difficult to distinguish.

6.2.1 *Gyroscope Data*

In order to address Hypothesis 2 regarding the ability of gyroscope data gathered at each keypress to distinguish between the three user positions of sit, stand and walk, we analyzed this data using two classifiers: C4.5 Decision Tree (DT) and Logistic Regression (LR). We used the Weka implementation of these classifiers [26], which were chosen because of their use in activity recognition and keystroke dynamics work, respectively. We used 10-fold cross validation as with the previous classifications.

	Pos.	Classifier Metric (%)		
DT		**FAR**	**FRR**	**AUC**
	Sit	4.5	10.5	97.3
	Stand	10.3	20.2	91.5
	Walk	9.2	23.3	92.2
LR		**FAR**	**FRR**	**AUC**
	Sit	10.8	18.6	90.8
	Stand	15.8	39.8	82.3
	Walk	17.7	31.1	84.5

Table 3: Gyroscope data FAR, FRR and AUC (%) results averaged over all participants for Decision Tree (DT) and Logistic Regression (LR) classifiers.

As can be seen in Figure 3 our results were promising for both DT AND LR, although slightly better for DT. AUC is a valuable measure of classifier accuracy for binary classification problems; Table 3 reports the AUC for the position in question considered the positive class, and the other two positions considered the negative class. For example, the AUC of 97.3% for the Sit position for DT is measured based on using Sit as the positive class and Stand and Walk together as the negative class. In general, values of greater than 90% for DT indicate that the gyroscope data gathered is very good at distinguishing between the three user positions. Note that the AUC values for both classifiers for the Sit position are higher than those values for Stand and Walk. We believe this is because users tended to prop their arms on a table while typing during the study, which may mean that the mobile device moved less (or at least differently) compared to the unsupported arm positions while in the Stand and Walk conditions. These promising results show support for accepting Hypothesis 2.

The FRR values in particular, though, are a bit worrisome as they are high for both classifiers. However, these results are not being used to determine authentication suitability, but only to justify using gyroscope data to determine user position. Thus, there is little security risk associated with misclassifying the user's position; such a misclassification simply means the wrong model may be used for classifying keystroke and bigram data. The selection of the wrong model may result in rejecting the legitimate user, which would require reauthentication and thus could affect usability. We intend to explore the impact of such misclassifications in future work.

6.2.2 *Keystrokes*

Once the user's position has been determined, key hold time and inter-key latency data from the user's typing patterns will be classified as owner or not-owner based on three trained models based on data from the three user positions of Sit, Stand and Walk. This section discusses the results of a feasibility study in which the study participants' keystroke and bigram data was classified using position-based models with the DT and LR classifiers to allow for easy comparison to the naïve results shown in Table 2.

Table 4 shows the FAR, FRR and AUC metrics that result from classifying key hold times over the three user positions. The results for DT for all three metrics are better than those for LR; FAR values for LR in the 18.7% to 20.42% range indicate an unacceptably high one in five chance that an attacker will be mistaken for the legitimate device owner. Furthermore, FRR values of about 23% for LR show a usability problem since nearly one in four authentication attempts by the legitimate owner will fail. Since keystroke dynamics is best used as secondary or continuous authentication method, such a high failure rate is not as great a problem as for primary authentication methods. However, it is still an unacceptably high reauthentication rate. Therefore, we intend to use DT as the classifier of choice in future work.

	Position	Classifier Metric (%)		
DT		**FAR**	**FRR**	**AUC**
	Sit	8.5	8.4	90.3
	Stand	8.3	9.3	89.8
	Walk	7.4	8.3	91.0
LR		**FAR**	**FRR**	**AUC**
	Sit	18.70	23.18	82.76
	Stand	19.63	23.73	82.32
	Walk	20.42	23.55	82.14

Table 4: Keystroke data (key hold time) FAR, FRR and AUC (%) results averaged over all participants for Decision Tree (DT) and Logistic Regression (LR) classifiers.

6.2.3 *Bigrams*

Previous studies have shown that bigrams on mobile devices are not distinctive as authenticators on mobile devices [40]. However, our results refute this result, perhaps due to the use of position as an initial classification. Table 5 shows that bigrams are, in fact, a quite accurate means of authentication. The table shows the results of classifying Sit, Stand and Walk data as separate classification problems; for example, the Sit row for each classifier shows the results of classifying only Sit data into Owner and Not Owner classes; similarly for the Stand and Walk rows.

Table 5 shows that the DT classifier outperforms the LR classifier for FRR results, while remaining only slightly higher than LR for FAR values. The AUC values show that inter-key latency is perhaps even slightly more distinctive than key hold time since the bigram AUC values are slightly higher than those of keystrokes. Given that our intent is to use keystroke dynamics for secondary or continuous authentication, AUC values of 89.82% to 93.61% for DT over the three positions are highly encouraging. As with the keystroke data results, we intend to use the DT classifier in future work since the AUC values are comparable to LR, but the FRR values for DT are considerably lower, indicating less likelihood of reauthentication, thereby supporting improved usability.

6.2.4 *Keystrokes + Bigrams*

Due to the encouraging keystroke and bigram results after

	Position	Classifier Metric (%)		
DT		**FAR**	**FRR**	**AUC**
	Sit	6.9	6.0	89.8
	Stand	6.6	6.9	93.6
	Walk	6.9	7.4	92.7
LR		**FAR**	**FRR**	**AUC**
	Sit	5.0	13.0	92.2
	Stand	4.3	12.7	93.1
	Walk	5.3	13.5	91.2

Table 5: Bigram data (inter-key latency) FAR, FRR and AUC (%) results averaged over all participants for Decision Tree (DT) and Logistic Regression (LR) classifiers.

position classification, we combined the key hold time and inter-key latency features while still classifying only one position at a time. Table 6 shows the results of this classification; as expected, combining features showed an increase in AUC for both classifiers, although the increase is more notable for the LR classifier. Furthermore, the FAR and FRR values from LR classification are lower for the combined features when compared to those features alone. AUC results of around 97% over all positions for LR move keystroke dynamics into a range we consider suitable for primary authentication, although this must be validated via simulation to determine the impact of battery and processor use, which we leave for future work. Thus, we recommend that keystroke dynamics be used only for secondary or continuous authentication.

	Position	Classifier Metric (%)		
DT		**FAR**	**FRR**	**AUC**
	Sit	5.6	6.1	93.2
	Stand	6.1	5.3	93.3
	Walk	4.8	5.6	93.9
LR		**FAR**	**FRR**	**AUC**
	Sit	1.7	7.0	97.3
	Stand	1.8	5.5	97.7
	Walk	1.4	6.2	97.7

Table 6: Combination of keystroke (key hold time) and bigram (inter-key latency) data FAR, FRR and AUC (%) results averaged over all participants.

The approximately 90% and up AUC values for DT over keystrokes, bigrams and their combination indicates that using keystroke dynamics as a distinctive information source for authentication is viable, and shows support for accepting Hypothesis 1 of this work.

6.2.5 *Comparison to Position Independent Results*

We now move to comparing the naïve, position independent keystroke and bigram results shown in Table 2 to the relevant data in Tables 4 and 5. The highest AUC for position independent results (Table 2) is 66.9% for key hold time data, and 60.7% for inter-key latency data, while the highest AUC values when position is considered are 91.01% for key hold time and 93.61% for inter-key latency. These increases are considerable, and show that considering position before authentication classification is a plausible approach to using keystroke dynamics as a secondary or continuous authentication method. This result is supported by the overall reduction in FAR and FRR values: from lows of 41.5% (FAR) and 23% (FRR) without considering position, to lows of 4.25% (FAR) and 6% (FRR) when position is considered. These reductions indicate that the two-phased approach is better able to minimize both attacker access and reauthentications compared to not considering position. These results show strong support for Hypothesis 3 regarding improvements in classification results when considering device position.

6.2.6 Implications

Our threat model outlined in Section 4 described possible attacks that can affect the system described in this paper. In particular, we stated that it is possible that the attacker may observe the device owner typing, and thus may be able to gather information that would allow the attacker to imitate the legitimate device owner. Given that the position-independent results showed us that a user's typing patterns are variable across positions, an attacker would have to learn different typing styles across all positions, which we consider unlikely.

The other implication to consider is what might happen if the first phase of the model (determining position) is incorrect. The effect would be that the wrong model would be used for matching the gathered keystroke information, which may result in rejecting a legitimate user. Given that our method is intended to be used for transparent authentication, there are two possibilities: either that multiple rejected authentication attempts are required to completely block access to the user, which enhances usability since additional user action is not required, but also has serious security ramifications since it increases the possible attack window. The second option is to prompt the device owner to enter a password or PIN when transparent methods are rejected, which has usability implications due to requiring additional user effort, but reduces the possible attack window. The preference for one of these options over the other depends on what type of system it is implemented in; a high-security system may require the latter.

7. LIMITATIONS

As with other user studies, ours has several limitations that must be considered in light of the results provided. Users often walked very slowly during the walking conditions; their focus was on their perceived goal (to enter the phrases) rather than on actually walking. It is likely that in a real-world situation, the user will be intent on walking rather than typing (i.e., if they are running late). Similarly, we observed users propping their arms on a table while typing during the Sit condition, and leaning against a wall during the Stand condition. It is possible that these postures introduced bias in that the static positional data may be more static, thereby further distinguishing this data from that gathered in the Walking condition. This may have resulted in better FAR, FRR and AUC values than in a real-world environment. The phrases themselves may have caused some bias in typing patterns (and removed some ecological validity) as the participant was transcribing the given phrases rather than creating true free text. Furthermore, our study required participants to use an unfamiliar mobile device with an unfamiliar keyboard, which may have had an effect on the participants' typing speed, as well as possibly changing how the keyboard reacts to touch events. We also disabled the predictive and corrective text actions, which affects ecological validity as these are widely used features on soft keyboards. We also did not consider hand postures during our study; participants were permitted to switch between typing with one thumb, both thumbs or any finger while in any of the six experimental conditions. We tested only a small set of classifiers (DT and LR) with few features. Many more possible classifiers exist, including those that take an anomaly detection approach, in which the classifier is trained only on the owner's data rather than adding in some representative negative samples. An anomaly detection approach is considered by some to be more valid for a single-user mobile device as it is unlikely that there will be a significant sample of other people's typing that can be used to create the negative class [10]. While other studies have achieved improved FAR and FRR values by using fused features in a multimodal biometric [10], we chose to use only inter-key latency and key hold time first to conform to other similar studies and also as a minimum baseline result to which future work can be compared. Finally, we collected data in a single session of only one hour in duration, which does not effectively study possible changes in a person's typing pattern over time.

Our final limitation is on the selection of Sit, Stand and Walk as user positions. We chose these based on our intuition that these are the most likely positions in which a user may type. It is unlikely, for instance, that a user will choose to (or be able to effectively) type while running, and positions such as laying down are very similar to both sitting and walking. We plan on addressing this issue with one of two approaches: either create an full activity recognition system that encompasses more positions, or narrowing the positions into those that are similar, such as ambulatory (e.g., walking, running) versus static (e.g., sitting, standing).

While each of these design decisions results in bias that will have differing effects on the results of this study, we believe that the largest effect will be in the overall classification rates, which in the worst case would be artificially high, which would give an inaccurate representation of the predictive power of gyroscope and keystroke data. We note that our results are similar to other studies in this field [39], and plan on removing some of these limitations (particularly those to do with the custom keyboard and disabling predictive and corrective text functions) in the simulation of our phased approach that we mention as future work.

8. CONCLUSION

In this paper we have presented the results of a user study designed to test the efficacy of keystroke dynamics as a potential continuous, transparent authenticator on mobile devices. We first determined via gyroscope data whether the typist was seated, standing or walking, then trained and tested three different models based on dynamic text from each of those three positions. We found that determining position first before classifying typing data resulted in an AUC increase of 30%. Our two-phased model approach of determining position first, then classifying keystroke information thus has merit and should be further examined via simulations. Both our experimental design and our threat model were chosen to provide as much ecological validity as possible given that the study was lab-based. We hope that taking a step back in assessing how much information is required per keystroke, and mimicking how users type in the wild, will provide an important advance in the field of keystroke dynamics.

Overall, our results support continuing research in keystroke dynamics as a transparent authenticator. We removed the need for a feature vector and the associated pre-processing required by them, while supporting a realistic evaluation scenario. We refuted previous results that showed bigram inter-key latency is not as distinctive as hoped for dynamic text, meaning that this feature may now be considered along with key hold time. We also provided support for the idea that transparent authentication may indeed be viable, which may help remove the need for a potentially intrusive and unusable authentication interface.

9. FUTURE WORK

We have begun creating a simulation of the phased approach pictured in Figure 1; we will use the simulation to test the effect of the phased approach on device battery and processor consumption, the amount of time needed for a classification decision, and the amount of data needed to reach suitable FAR, FRR and AUC values for continuous authentication. The use of a simulation as a first step will allow us to more closely model real-world typing conditions since our results were from a lab study. With the results of the simulation as a guide, we also plan on creating a prototype of this authentication method for Android devices, which we will test via a longitudinal user study. We will also use the simulation to innovate solutions to the sit-stand confusion, as well as to determine whether a catch-all classifier is suitable for situations where the user is neither sitting, standing nor walking while typing.

Acknowledgements

The authors wish to thank Thomas C. Eskridge, Philip Bernhard, and Abdullah Alharbi for valuable discussions that shaped this work.

10. REFERENCES

[1] A. A. E. Ahmed, I. Traore, and A. Almulhem. Digital Fingerprinting Based on Keystroke Dynamics. In *Proceedings of the Second International Symposium on Human Aspects of Information Security & Assurance (HAISA 2008)*, pages 94 – 104, July 2008.

[2] J. M. Allen, L. A. McFarlin, and T. Green. An In-Depth Look into the Text Entry User Experience on the iPhone. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 52(5), pages 508 – 512, 2008.

[3] A. Alsultan and K. Warwick. Keystroke Dynamics Authentication: A Survey of Free-text Methods. *International Journal of Computer Science Issues*, 10(4), 2013.

[4] A. Alsultan and K. Warwick. User-Friendly Free-Text Keystroke Dynamics Authentication for Practical Applications. In *IEEE International Conference on Systems, Man and Cybernetics*, pages 4658 – 4663, 2013.

[5] A. Anjum and M. U. Ilyas. Activity Recognition Using Smartphone Sensors. In *Proceedings of First Workshop on People Centric Sensing and Communications*, pages 914 – 919, 2013.

[6] S. Azenkot and S. Zhai. Touch Behavior with Different Postures on Soft Smartphone Keyboards. In *Proceedings of 14th International Conference on Human Computer Interaction with Mobile Devices and Services*, pages 251 – 260, 2012.

[7] F. Bergadano, D. Gunetti, and C. Picardi. User Authentication Through Keystroke Dynamics. *ACM Transactions on Information and System Security*, 5(4):367–397, November 2002.

[8] T. Brezmes, J.-L. Gorricho, and J. Cotrina. Activity Recognition from Accelerometer Data on a Mobile Phone. In *10th International Work-Conference on Artificial Neural Networks (IWANN)*, volume 5518 of *Lecture Notes in Computer Science*, pages 796 – 799, 2009.

[9] A. Buchoux and N. Clarke. Deployment of Keystroke Analysis on a Smartphone. In *Proceedings of the 6th Australian Information Security Management Conference*, pages 40 – 47, 2008.

[10] D. Buschek, A. D. Luca, and F. Alt. Improving Accuracy, Applicability and Usability of Keystroke Biometrics on Mobile Touchscreen Devices. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems (CHI)*, page to appear., 2015.

[11] T.-Y. Chang, C.-J. Tsai, and J.-H. Lin. A Graphical-Based Password Keystroke Dynamic Authentication System for Touch Screen Handheld Mobile Devices. *Journal of Systems and Software*, 85(5):1157 – 1165, May 2012.

[12] T. Chen, Y. Yesilada, and S. Harper. What Input Errors do you Experience? Typing and Pointing Errors of Mobile Web Users. *International Journal of Human-Computer Studies*, 68(3):121–182, 2010.

[13] N. Clarke and S. Furnell. Advanced User Authentication for Mobile Devices. *Computers & Security*, 26(2):109 – 119, March 2007.

[14] N. Clarke and S. Furnell. Authenticating Mobile Phone Users Using Keystroke Analysis. *International Journal of Information Security*, 6(1):1 – 14, January 2007.

[15] N. Clarke, S. Karatzouni, and S. Furnell. Transparent Facial Recognition for Mobile Devices. In *Proceedings of the 7th International Information Security Conference*, 2008.

[16] J. Clawson, T. Starner, D. Kohlsdorf, D. P. Quigley, and S. Gilliland. Texting While Walking: An Evaluation of Mini-QWERTY Text Input While On-the-Go. In *Proceedings of 16th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 339 – 348, 2014.

[17] M. Conti, I. Zachia-Zlatea, and B. Crispo. Mind How You Answer Me!: Transparently Authenticating the User of a Smartphone when Answering or Placing a Call. In *Proceedings of the 6th ACM Symposium on Information, Computer, and Communications Security*, pages 249 – 259, 2011.

[18] H. Crawford. Keystroke Dynamics: Characteristics and Opportunities. In *Proceedings of 8th Annual International Conference on Privacy, Security and Trust*, pages 205 – 212, 2010.

[19] H. Crawford, K. Renaud, and T. Storer. A Framework for Continuous, Transparent Mobile Device

Authentication. *Computers & Security*, Vol. 39, Part B:127 – 136, 2013.

[20] M. Frank, R. Biedert, E. Ma, I. Martinovic, and D. Song. Touchalytics: On the Applicability of Touchscreen Input as Behavioral Biometric for Continuous Authentication. In *IEEE Transactions on Information Forensics and Security*, volume 8, pages 136 – 148, 2012.

[21] C. Giuffrida, K. Majdanik, Mauro, and H. Bos. I Sensed It Was You: Authenticating Mobile Users with Sensor-Enhanced Keystroke Dynamics. In *Proceedings of Detection of Intrusions and Malware, and Vulnerability Assessment*, volume 8550 of *Lecture Notes in Computer Science*, pages 92 – 111, 2014.

[22] M. Goel, L. Findlater, and J. Wobbrock. WalkType: Using Accelerometer Data to Accomodate Situational Impairments in Mobile Touch Screen Text Entry. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*, pages 2687 – 2696, 2012.

[23] D. Gunetti and C. Picardi. Keystroke Analysis of Free Text. *ACM Transactions on Information and System Security*, 8(3):312 – 347, August 2005.

[24] Q. Guo, B. Liu, and C. W. Chen. A Two-Layer and Multi-Strategy Framework for Human Activity Recognition Using Smartphone. In *IEEE International Conference on Communications (ICC)*, 2016.

[25] P. Gupta and T. Dallas. Feature Selection and Activity Recognition System Using a Single Triaxial Accelerometer. *IEEE Transactions on Biomedical Engineering*, 61(6):1780 – 1786, 2014.

[26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1), 2009.

[27] K. S. Killourhy and R. A. Maxion. The Effect of Clock Resolution on Keystroke Dynamics. In R. Lippmann, E. Kirda, and A. Trachtenberg, editors, *Proceedings of RAID 2008*, volume 5230 of *Lecture Notes in Computer Science*, pages 331–350. Springer-Verlag Berlin Heidelberg, 2008.

[28] K. S. Killourhy and R. A. Maxion. Comparing Anomaly-Detection Algorithms for Keystroke Dynamics. In *Proceedings of the IEEE/IFIP International Conference on Dependable Systems & Networks*, pages 125 – 134, 2009.

[29] P. O. Kristensson and K. Vertanen. Performance Comparisons of Phrase Sets and Presentation Styles for Text Entry Evaluations. In *Proceedings of 12th ACM International Conference on Intelligent User Interfaces*, pages 29 – 32, 2013.

[30] J. Leggett, G. Williams, M. Usnick, and M. Longnecker. Dynamic Identity Verification via Keystroke Characteristics. *International Journal of Man-Machine Studies*, 35(6):859–870, December 1991.

[31] I. S. MacKenzie and R. W. Soukoreff. Phrase Sets for Evaluating Text Entry Techniques. In *Extended Abstracts on Human Factors in Computing Systems*, pages 754 – 755, 2003.

[32] E. Maiorana, P. Campisi, N. Gonzalez-Carballo, and A. Neri. Keystroke Dynamics Authentication for Mobile Phones. In *Proceedings of the 2011 Symposium on Applied Computing*, pages 21 – 26, 2011.

[33] P. Marquardt, A. Verma, H. Carter, and P. Traynor. (sp)iPhone: Decoding Vibrations from Nearby Keyboards Using Mobile Phone Accelerometers. In *Proceedings of ACM Conference on Computer and Communication Security*, pages 551 – 562, 2011.

[34] A. Messerman, T. Mustafic, S. A. Camtepe, and S. Albayrak. Continuous and Non-intrusive Identity Verification in Real-Time Environments Based on Free-Text Keystroke Dynamics. In *2011 International Joint Conference on Biometrics*, pages 1 – 8, 2011.

[35] F. Monrose, M. Reiter, and S. Wetzel. Password Hardening Based on Keystroke Dynamics. *International Journal of Information Security*, 1(2):69 – 83, February 2002.

[36] F. Monrose and A. D. Rubin. Keystroke Dynamics as a Biometric for Authentication. *Future Generation Computer Systems*, 16:351–359, 2000.

[37] M. F. M. Noor, S. Rogers, and J. Williamson. Detecting Swipe Errors on Touchscreens using Grip Modulation. In *Proceedings of Conference on Human Factors in Computing Systems (CHI)*, pages 1909 – 1920, 2016.

[38] Paul Dunphy and Andreas P. Heiner and N. Asokan. A Closer Look at Recognition-Based Graphical Passwords on Mobile Devices. In *Proceedings of the 6th Symposium on Usable Privacy and Security*, pages 26 – 38, 2010.

[39] C. Shen, T. Yu, S. Yuan, Y. Li, and X. Guan. Performance Analysis of Motion-Sensor Behavior for User Authentication on Smartphones. *Sensors*, 16(3), 2016.

[40] T. Sim and R. Janakiraman. Are Digraphs Good for Free-Text Keystroke Dynamics? In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1 – 6, 2007.

[41] X. Suo. A Study of Graphical Password for Mobile Devices. In *Proceedings of 5th International Conference on Mobile Computing, Applications and Services*, pages 202 – 214, 2014.

[42] H. Xu, Y. Zhou, and M. R. Lyu. Towards Continuous and Passive Authentication via Touch Biometrics: An Experimental Study on Smartphones. In *Proceedings of Symposium on Usable Privacy and Security*, 2014.

Impact of User Characteristics on Attitudes Towards Automatic Mobile Application Updates

Arunesh Mathur
University of Maryland
College Park, MD
amathur@umd.edu

Marshini Chetty
Princeton University
Princeton, NJ
marshini@princeton.edu

ABSTRACT

To keep mobile devices secure, experts recommend turning on auto-updates for applications, but recent research has suggested that users often avoid auto-updating because updates can lead to undesirable consequences such as user interface changes or compatibility issues. Understanding whether there are commonalities amongst users who avoid auto-updates can help us create better mobile application updating interfaces. However, little is known about how users' characteristics associate with their attitudes towards auto-updating their mobile applications, or how we can leverage these characteristics to encourage users to auto-update these applications to improve security. In this paper, by surveying Android users, we establish how users' past experiences with software updating, and users' psychometric traits differentiate those users who avoid application auto-updates from those who do them, as well as users' preferences towards auto-updating their applications. Our findings reveal that users who avoid application auto-updates are more likely to have had past negative experiences with software updating, tend to take fewer risks, and display greater proactive security awareness. Users' perceived level of trust with mobile applications also determined how comfortable they are auto-updating these applications. Based on these findings, we recommend how Android can improve the design of application update systems to encourage users to auto-update and keep their devices secure.

1. INTRODUCTION

Keeping mobile applications and platforms updated is important since these devices store sensitive data from or about users and software updates can prevent security exploits from known vulnerabilities. For instance, in 2015 alone, Symantec reported 528 new mobile vulnerabilities [1], up 214% from 2014. Furthermore, recent research has shown that mobile users are slow to apply updates: only 32% users auto-updated their applications and only 16% applied updates manually as soon as updates were released [2], and only half of all users update to a new application version within the first week after the update's release [3]. Additionally, on the Android Operating System (OS)—the mobile OS with the largest market share of mobile phones worldwide [4]—studies [5, 6] have reported multiple Secure Sockets Layer (SSL) and OpenSSL Heartbleed bug [7] vulnerabilities which could have been fixed by promptly applying application updates. For this reason, various companies [8, 1], academics [9], and even the United States-Computer Emergency Readiness Team [10] suggest that developers deploy, and users turn on automatic updates—updates that are downloaded and installed without human intervention—to ensure that their systems remain secure. Automated updates have also been shown to be effective, more so than requiring users to manually update their devices [9, 11].

However, recent research has suggested that users often turn off automatic updates since updates can disrupt settings, cause unnecessary reboots, compatibility issues, or change the user interface [12, 13]. Yet we know little about whether there are commonalities amongst those users who avoid auto-updates versus those who do auto-update, despite knowing that user characteristics can influence computer security attitudes and behaviors [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24]. To address this gap in the literature and determine if we can leverage user characteristics and auto-updating preferences to encourage more users to auto-update their mobile applications, we posed two research questions. First, we asked whether there are underlying user characteristics that differentiate users who avoid auto-updating from those who auto-update their Android applications based on their current auto-update settings. Second, we asked how these user characteristics, including users' attitudes towards their Android applications explain users' preferences indicating whether they would like auto-updating to vary across their applications. Our goal is to inform the design of enhanced mobile application systems that increase user uptake of auto-updates on mobile devices to improve mobile security.

To answer these questions, we conducted a survey with 477 Android users on Amazon Mechanical Turk in the United States (US). Specifically, we considered how users' past experiences with software updating, users' psychometric traits—including their risk taking capacities, consideration for future consequences, propensity to engage in cognitive endeavors, and resistance to change—differentiate how users currently auto-update their Android applications. Next, we investigated how these user characteristics, including users' attitudes towards their Android applications—such as how much they trust an application—explain how comfortable users

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

are auto-updating updates across their different applications, that is, their auto-updating preferences.

In a new contribution, our results show that users who avoid auto-updating their applications on Android differ from those who auto-update in three ways:

1. First, we found a large effect indicating that users who avoid auto-updating have had previous negative experiences with updating their software—confirming the findings from previous interview studies about desktop users [12, 13]. However, unlike prior work, our findings suggest users' past negative experiences with updates may not necessarily have been on their Android devices.

2. Second, we found a medium effect indicating that users who avoid auto-updating also tend to take fewer financial investment risks. For instance, these users were less likely to invest their money in business ventures and mutual funds. In addition, we found a medium effect indicating that users who avoid auto-updating also tend to fewer ethical risks. For instance, these users were less likely to take questionable deductions on their income task returns (to receive greater returns).

3. Third, we found a small effect indicating that users who avoid auto-updating also exhibit a greater propensity to take proactive steps to maintain their online security—similar to what others have found for users who avoid auto-updating on desktops [13]. For instance, these users were more likely to verify the green HTTPS lock on websites, and verify links before clicking them.

Finally, we discovered medium effects indicating that users were less comfortable auto-updating across their Android applications if they had a previous negative experiences with software updating, whereas users were more comfortable auto-updating across their Android applications for security updates and when they perceived an application as trustworthy—similar to what others studies have reported for desktop users [25].

Based on our findings, we make four primary recommendations to improve the design of mobile application updates on Android to encourage users to auto-update. First, we suggest that an improvement to the current Android OS would be to provide users with a more accessible mechanism to rollback application updates to a prior point in time to encourage users to be more risk taking with respect to turning on auto-updates. Second, we suggest leveraging the characteristics we identified of users who avoid auto-updating, including their risk averse nature, to design nudges and messages to encourage users into auto-updating security updates. Third, we suggest that the security community study the practices of software developers, how they develop and build updates, and how these practices lead to negative experiences for end-users. Finally, we suggest that an improved Android application interface for updates could be personalized by inferring users' attitudes towards their Android applications and preferences for auto-updating those applications using our work as a starting point. Doing so, may encourage more users to auto-update their mobile applications, which in-turn will ultimately affect the security of their mobile devices. In the remainder of this paper, we discuss related work, our methods and study, our findings, discussion, and conclusions for improving mobile software update interfaces.

2. BACKGROUND AND RELATED WORK

In this section, we highlight previous research related to software updates, and place our research in context.

2.1 User Characteristics and Security

Multiple studies have investigated how users' individual differences affect their security attitudes and behaviors. One such line of work [21, 22, 23, 24] examined how demographics and users' level of technical expertise impact security decisions. For instance, Jeske *et al.* [21] suggested that the manner in which users select a wireless network may be affected by individual differences in users. These authors designed and evaluated user interfaces in a 67 participant study showing how certain interface elements (e.g., color and ordering) can be utilized to help users with low technical expertise select secure networks. Another set of researchers [24] examined the characteristics of users who succumb to phishing attacks, and found that younger users and women were more susceptible than other users. These researchers designed and evaluated educational material to help these users stay protected against phishing attacks. In another study of targeted security solutions, Garg *et al.* [22] found that using video to communicate malware and phishing threats improved older adults' ability to understand security risks with these two threats.

Another line of work [14, 15, 16, 17, 18, 19, 20] examined how differences in users' beliefs, mental models and decision making capacities impact security attitudes and behaviors. Wash [14] uncovered how users' mental models and beliefs about computer security influence the actions they take to protect themselves across two folk models of threats: hackers and viruses. The author suggested that due to the variance in users' beliefs about security threats, one-size-fits-all security interfaces such as warnings and notifications may be less impactful than those that are specifically targeted at users' beliefs. Whitty *et al.* [15] found that older people and people who self-regulated their actions and behavior—as measured by the self-monitoring personality trait—were more likely to share their passwords with others. In related work, Egelman and Peer developed [16] and validated [26] the Security Behavior Intentions Scale (SeBIS) to measure behavior intentions, and with their data, demonstrated [17] that these behavior intentions are associated with users' psychometric traits, including capacity to take risks, being inherently curious and inquisitive, and thinking about long-term implications of actions. For instance, the authors discovered that those users who take fewer risks also tend to keep their software updated to prevent potential harm from exploits. The authors suggested that segmenting users by these traits may allow designers to infer users' security intentions, and use them to tailor computer security user interfaces to help users remain secure. Similarly, in another study, Malkin *et al.* [20] developed and tested browser SSL warnings tailored to users' decision making capabilities, and found several correlations between the framing of the warnings and users' decision making capabilities. Related studies have also shown that users who take fewer risks were also less likely to plug in potentially harmful USB drives [19] and more likely to keep their systems secure [18]. Our study builds upon both lines of work on linking users' past software update experiences, their psychometric traits, and their security behavior (which is driven by underlying beliefs about security) to attitudes towards automatic application updates.

2.2 Software Updates

2.2.1 User Issues Around Software Updates

There is a growing body of work that examines in detail users' attitudes and interactions with software updates but most of this work focuses on the desktop experience of updates. One set of studies has examined how users manage their computer security, perceive software updates, and software updating behaviors. For example, Ion *et al.* [27] compared the security advice expert and non-expert users gave to others to stay secure, and found that non-experts lacked awareness about the benefits of software updates and used their judgements to avoid updates that they felt introduced bugs. Furthermore, they found that 39% of experts reported auto-updating compared to 29% of non-expert users. In another US national representative survey [28], researchers found that a large fraction of their sample updated their software with 39% reporting they update their software immediately and 41% reporting that they update their software sometime after an update is released, and only 5% reported rarely or never updating their software.

Several studies have indicated that negative experiences—such as user interface changes or compatibility issues with software—either cause users to avoid or delay software updates [25, 12, 29, 13] on desktop machines. In some cases, users avoid desktop updates because they find software updates messages confusing and unclear [30]. In other cases, studies [31, 32] show that users often delay and only perform updates on Wi-Fi networks if they have access to limited and expensive Internet data plans on both desktop and mobile.

There are at least three studies of mobile updating behaviors, specifically on the Android OS. Moller *et al.* [33] and Oltrogge *et al.* [3] found that half of the users they studied would still use a vulnerable application on their phones even seven days after the release of the update that fixed that vulnerability. Tian *et al.* [34] developed a novel updating notification that used user generated reviews to help mobile users make privacy conscious decisions about which updates to apply based on what permissions were asked for by the updates. Collectively, all of these studies illuminate that users lack awareness about the security benefits of software updates, and that users often delay or avoid updates. However, unlike our work, these studies do not focus specifically on mobile users' automatic update experiences or link auto-updating attitudes and preferences with user characteristics.

2.2.2 Automatic Software Updating

Several studies have shown that auto-updating is extremely effective at keeping users up to date with the latest security patches. For instance, Gkantsidis *et al.* [35] analyzed software update data from close to 300 million Microsoft Windows computers and discovered that more than 90% of all machines which had automatic updates enabled had security patches applied. Similarly, Duebendorfer and Frei *et al.* [11] collected log and update data from various Web browsers including Google Chrome, Mozilla Firefox, and Apple Safari. They found that compared to all other browsers, Google Chrome's silent update mechanism, a form of automatic updates that requires no user notification, had the most efficient patching rate: within three weeks, 97% of all active users were up-to-date on the latest version, unlike browsers with other update mechanisms such as Mozilla Firefox and Apple Safari. One other study of nearly 8.4 million hosts [9] also demonstrated that applications with auto-updating mechanisms such as Chrome reached 50% and 90% patch deployment coverage significantly faster than those that did not, such as Wireshark.

Another set of studies touches more on user experiences with automatic updates. Two studies about desktop users revealed that automatic updates can lead to varying user experiences. First, this research discovered that because automatic updates do not include users in the decision making process, users develop poor mental models of how updates on their system work [36]. As a result of these poor mental models, the authors argue that users fail to troubleshoot and manage these updates, which adversely affects the security of their systems [37]. Second, in another study, researchers found that users who desire control and make active choices in computer security and maintenance tasks turned off automatic updates, and used their own judgement to decide which updates to apply—but were sometimes less secure than those who kept auto-updating on [13]. Finally, Mathur *et al.* [25] designed a novel interface to support silent updates and found that users varied in their preferences to let applications auto-update with some preferring the convenience of auto-updates and others disliking the lack of control over what changes updates make to their systems.

While these studies collectively suggest that automatic updates are indeed effective to patch systems, users are still impacted when applications and devices auto-update. These studies also highlight several qualitative user experiences around auto-updates on desktops, but they offer no such insights into mobile users' attitudes towards auto-updating. Our study makes the following contributions to the body of work on users and updates: we provide evidence of differences between users who avoid and who do auto-update their applications on mobile devices, we show what factors explain user preferences for mobile application auto-updates, and we make recommendations for leveraging this information to design better mobile update interfaces to increase the chance of users' auto-updating to remain secure.

2.3 Android Application Updates

Given that it has the large market share of all OSes that run on mobile devices [4], we decided to study application software updating on the Android OS. Android users rely on the Google Play Store to download and update applications on their phone [38]. The Play Store contains settings that allow users to control how they receive updates to their applications. As shown in Figure 1, these settings are:

1. **Do not auto-update apps:** Applications are not auto-updated, and users receive notifications each time updates become available for their applications.
2. **Auto-update apps at any time. Data charges may apply:** Applications are auto-updated without user consent regardless of whether the user is on Wi-Fi or on mobile data.
3. **Auto-update apps over Wi-Fi only:** Applications are auto-updated without user consent but on Wi-Fi only to prevent excessive data charges.

By default, the Android OS ships with "Auto-update apps over Wi-Fi only" option enabled. In addition to these options, Android allows users to disallow auto-updating certain applications even in auto-update mode in case they wish to provide consent to updating these applications. For instance, if users wish to provide consent to update the Google Chrome application, they can retain the default auto-update setting but disable auto-updating specifically for Google Chrome. Users can also roll back application updates for certain applications and not for others. Specifically, applications that come pre-installed with the device can be rolled back only to their initial version from within the Android OS *Settings* menu [39], whereas applications downloaded from the Play Store cannot be rolled back at all.

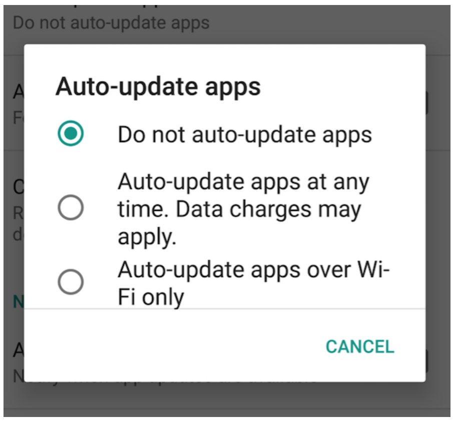


Figure 1: The Application Update Options Available in the Google Play Store on the Android OS.

If Android users disable auto-updating their applications, they see an update notification indicating the applications with available updates. Upon clicking the notification, users interact with a list of applications requiring updates. Users also see a notification—indicating the updated applications—after applications have been updated, independent of the update mechanism. All of these notifications can be activated or deactivated from the Play Store. Updates that require additional permissions for an application cannot be automated to prevent malicious applications from acquiring device permissions. However, since the introduction of Android 6.0, permissions are requested at run-time [40].

3. RESEARCH QUESTIONS

In order to design and develop better user interfaces and systems that encourage users to auto-update, we need to answer two questions. First, we need to identify what differences exist between the characteristics of those users who currently avoid auto-updating and those users who currently auto-update their mobile applications. Identifying these differences can help inform how the user interfaces of mobile devices can be improved to better incorporate these differences. Second, we need to identify—regardless of whether users currently auto-update—how these characteristics explain users' preferences indicating whether they would like auto-updating across their different applications. Establishing which factors explain these preferences can further help us identify how auto-update systems can incorporate these preferences into their design. In this section, we describe these research questions in further detail.

3.1 How are Users Who Avoid Auto-updating Different From Those Who Auto-update?

Our first research question investigated how various user characteristics differentiate those Android users who avoid auto-updating their applications from those who auto-update their applications (retaining the default option to auto-update applications in the Play Store). Previous interview-based studies have suggested that past negative experiences with software updating—such as surprise user interface changes or compatibility issues—can affect users' auto-updating behavior [12, 13] and their attitudes towards future updates [29]. Based on this observation, we formulated our first hypothesis:

• **H1**: Avoiding auto-updates will likely be associated with users who have had negative experiences with updating their software.

Previous research [17, 16] into users' psychometric traits—including risk taking capacities, consideration for future consequences, propensity to engage in cognitive endeavors—and cybersecurity behaviors has shown that they correlate with how often users' take actions and make decisions towards keeping their software updated. Unlike when users are asked to provide consent to updating each time an update is available, auto-updating is a one-time decision users make to allow their system to update itself and does not require continual consent. However because auto-updates are installed without users' consent, they may cause undesired consequences, and are likely to be avoided by those who take fewer risks. Based on this, we formulated our second hypothesis:

• **H2**: Avoiding auto-updates will likely be associated with lower risk taking behavior.

Next, leaving auto-updates on has the potential to cause undesired consequences to systems in the long-term, and are therefore likely to be avoided by those who consider the future consequences of their actions. This allowed us to formulate our third hypothesis:

• **H3**: Avoiding auto-updates will likely be associated with a higher consideration of future consequences.

When updates are automatically installed, users only grasp the changes made by the update after the update has been installed. Therefore, it is likely that those who avoid auto-updates also have a greater propensity to keep appraised of the changes updates make. This led to our fourth hypothesis:

• **H4**: Avoiding auto-updates will likely be associated with higher curiosity and inquisitiveness.

Finally, because auto-updating applies updates as soon as the updates become available and updates have the potential to bring about undesired changes and consequences, users who avoid auto-updating may also exhibit a greater resistance to change. This led to our fifth and final hypothesis:

• **H5**: Avoiding auto-updates will likely be associated with a greater resistance to change.

3.2 How Do User Characteristics Explain Users' Auto-updating Preferences?

Our second research question explored how user characteristics explain Android users' preferences towards auto-updating their mobile applications. Previous research [25] with desktop users has suggested that users vary in their preferences towards auto-updating their applications, and that they consider a variety of factors towards deciding which application to auto-update. The study suggested that users are more comfortable auto-updating security updates compared to non-security updates and applications they trust, and that they are less comfortable auto-updating applications that are important to them, applications they use frequently, and applications they are satisfied with. While we do not claim that these are the only characteristics Android users consider, we investigate how each of they factors correlate with users' auto-updating preferences across their applications in the mobile space. Therefore, to answer our second research question, we first considered how comfortable would users be auto-updating each application if they were given the choice to selectively auto-update their applications. Following that, we investigated how users' past negative experiences with software updating, users' psychometric traits, and users' attitudes towards auto-updating their Android applications explain these preferences.

4. METHOD

To answer our research questions we conducted a survey of Android users on Amazon Mechanical Turk (AMT) between April and May 2016. In total, the survey took took approximately 15 minutes to complete. We hosted the survey on SurveyGizmo[1], and advertised it as an "Android Apps Update Survey" task on AMT. Turkers were invited to participate if they were 18 or older, their primary smartphone was an Android phone, and if they had previously used the Play Store for at least one month. We used these filters to ensure participants were familiar with the Android OS and how applications are installed and updated. To ensure response quality, we restricted the task to Turkers based in the US who had an approval rating of 95% or higher. Because we did not filter Turkers based on the number of tasks previously completed, we added three attention check questions to the survey—based on the findings of Peer *et al.* [41]—one of which specifically asked about the Android OS. We filtered all responses that failed any of the attention check questions, and compensated the Turkers with $2.50 for completing the survey. The study was approved by the Institutional Review Board of our university.

4.1 Survey Instrument

The survey instrument contained three sections in total—all of which are described below—and is available in the Appendix.

4.1.1 Section One: Psychometric Scales

To answer our first research question, investigating what differentiates those who avoid auto-updates for their applications from the other users, we employed a psychometric scale to measure the corresponding psychometric trait for each hypotheses listed in Section 3. The statements of each scale and the order of the scales themselves were randomized to avoid any bias, and therefore each scale appeared in a page by itself in the survey. Specifically, we used the following psychometric scales, which have been used by Egelman and Peer [16, 17] and other studies in the past [18, 19, 24]:

Risk Taking: For hypothesis **H2**, we employed the Domain Specific Risk Taking scale (DoSpeRT) [42] to measure people's risk taking propensity. The DoSpeRT scale measures risk taking across the following dimensions: Ethical (e.g., passing off somebody else's work as your own), Financial/Investment (e.g., investing 10% of your annual income in a moderate growth mutual fund), Financial/Gambling (e.g., betting a day's income at the horse races), Social (e.g., admitting that your tastes are different from those of a friend), Recreational (e.g., bungee jumping off a tall bridge), and Health & Safety (e.g., engaging in unprotected sex). The score for each sub-scale lies between 1 (Extremely Unlikely) and 7 (Extremely Likely).

Future Consequences: For hypothesis **H3**, we employed the Consideration for Future Consequences scale (CFC) [43], which measures how much people consider the long-term consequences of their actions and decisions (e.g., my behavior is generally influenced by future consequences). The score for this scale lies between 1 (Extremely Uncharacteristic of Me) and 7 (Extremely Characteristic of Me).

Cognitive Endeavors: For hypothesis **H4**, we employed the Need for Cognition scale (NFC) [44], which measures how much people consider and indulge in thought and curiosity provoking endeavors (e.g., when I make a decision, I think about how it might affect me in the future). The score for this scale lies between 1 (Extremely Uncharacteristic of Me) and 5 (Extremely Characteristic of Me).

Resistance to Change: For hypothesis **H5**, we employed the Resistance to Change scale (RTC) [45], which measures how averse users are to change across the following dimensions: Short-term Focus (the extent to which individuals are distracted by the short-term inconveniences associated with change; e.g., often, I feel a bit uncomfortable even about changes that may potentially improve my life), Emotional Reaction (the amount of stress and uneasiness induced by change; e.g., when I am informed of a change of plans, I tense up a bit), Routine Seeking (inclination to adopt routines; e.g., whenever my life forms a stable routine, I look for ways to change it), and Cognitive Rigidity (frequency and ease with which individuals change their minds; e.g., my views are very consistent over time). The score for each sub-scale lies between 1 (Strongly Disagree) and 6 (Strongly Agree).

In addition, we also included the SeBIS scale to determine whether users who avoid auto-updating and do auto-update their applications differed in their security behavior intentions. That is, we wanted to determine if applying auto-updates is associated with a tendency to indulge in "good" security behaviors? We did so to determine whether users who intend to behave securely also consider auto-updating a "good" security practice. The SeBIS scale measures users' security behavior intentions across four dimensions: Password Generation (do users create strong passwords?; e.g., when I create a new online account, I try to use a password that goes beyond the site's minimum requirements), Proactive Awareness (do users take proactive steps towards their security?; e.g., when browsing websites, I mouseover links to see where they go, before clicking them), Device Updating (do users update their software and devices regularly?; e.g., when I'm prompted about a software update, I install it right away), and Device Securement (do users protect their devices with passwords and pins?; e.g., I use a PIN or passcode to unlock my mobile phone). The score for each sub-scale lies between 1 (Never) and 5 (Always).

[1] http://surveygizmo.com

4.1.2 Section Two: Android Application Update Settings and Auto-Update Preferences

Section Two of the survey collected data about participants' mobile application update settings and their preferences towards auto-updating their applications on Android. To understand how users currently auto-update their applications, we collected users' Android application update settings—as described in Section 2.3—in the Play Store. Rather than asking participants directly about whether they auto-updated applications on their devices, we elicited their auto-update settings in the Play Store using detailed labelled instructions to increase the validity of the self-reported data. More specifically, we asked users to report two settings: the first question asked participants to report their application update settings from the Play Store—one of the options in Figure 1—and the second question followed up with an image displaying how an application could be disallowed from auto-updating, asking participants if they ever disallowed auto-updates for any application. The second question appeared only if participants reported auto-updating in the first question.

To answer our second research question and elicit users' auto-updating preferences for mobile applications, we asked the participants to select the applications installed on their phones from a visual grid containing the most installed Android applications of all time [46]—all 95 of them—since more users are likely to encounter these applications. To eliminate any biases, the order of the applications in the grid was randomized for each participant. For each application, participants then rated their level of comfort with automating security and non-security updates (UpdateType)—assuming no data charges applied—on a scale of Uncomfortable (0) to Comfortable (100) using visual slider. Participants then answered, for each application, how trustworthy they felt the application was (Trust), how important the application was to them (Importance), how frequently they used the application (Use Frequency), and how satisfied they were with the application (Satisfied)—all on a scale of 1 (Lowest) to 5 (Highest). To limit the number of applications participants answered these questions about and prevent fatigue, we randomly drew a maximum of 10 applications from the ones participants reported they had on their phone. We designed the survey to elicit preferences in this manner—for each application—to gather how participants varied in their responses. To prevent ordering bias, we randomized how participants observed Sections One and Two of the survey.

4.1.3 Section Three: Past Update Experiences

Finally, in Section Three of the survey, we asked participants whether they had previously had a negative experience with software updates i.e., if they had regretted updating any device or software—not just their mobile devices. If participants answered "Yes", we listed the common negative experiences users reported with software updating from the literature at the time [13, 29, 25] for participants to select or to enter their own experiences. This section of the survey always appeared right before the end of the survey—to avoid any priming effects of asking participants to report their preferences towards auto-updating in Section Two. The survey ended with demographic questions asking about age, gender, income and occupation.

4.2 Survey Pilot and Deployment

After constructing the survey, we piloted it with a six participant convenience sample drawn from within our institution. During the pilot, we employed cognitive interviews [47], a commonly used technique in survey research, where participants are asked to "think-aloud" when attempting the survey questions. These are conducted to ensure the survey questions convey their intended meanings and measure the construct the researcher intends to measure. During the interviews, participants were asked to describe, in their own words, what the purpose of each question was, and whether they experienced any difficulty in answering them; we also specifically focused on the instructions to report the auto-update settings to ensure they were easy to follow through. Subsequently, we used these results to refine and revise a few questions. These interviews lasted for about 30 minutes. Participants were not reimbursed for their time.

4.3 Limitations of Self-Reported Data

We asked participants to self-report their Android application update settings in the survey, and therefore this data on settings could be subject to error. To limit this error, we asked participants to follow a set of well-labelled instructions to find and report the current settings on their phones instead of asking them directly about how they updated their applications. However, it is still possible that participants may have responded by recollecting or guessing their update settings as opposed to actually checking their devices. We did consider asking participants to upload a screenshot of the update settings page from the Google Play Store application but we chose not to do so since this would require participants to upload a screenshot from their mobile device to the device they were using to taking the survey, which may have been different. This additional step would have been prohibitively time consuming and cumbersome, and likely to have introduced a bias of its own. Future research could identify ways to collect this information directly from users' devices to increase the accuracy of the settings data.

4.4 Participants

The survey received 525 responses in total from AMT, out of which 48 responses (9.1%) failed at least one attention check question and were subsequently discarded, leaving 477 valid responses. Table 1 summarizes the survey participants' demographics. The survey participants were predominantly between 18–34 years old, with more males than females (62.3% vs 37.1% respectively). Nearly 90% of the survey participants had either a college or bachelor's degree, and earned a median annual income between $35,000 and $49,999. 33% of the survey participants reported that they avoided auto-updating their applications, i.e., they chose the "Do not auto-update apps" option described in Section 2.3 in the Play Store. Only close to 10% of the participants reported that they restricted auto-updating for at least one application using the feature described in Section 2.3.

4.5 Data Analysis

4.5.1 Reliability and Dimensionality of the Scales

Before proceeding towards building statistical models for our data, we analyzed all the psychometric data in terms of *Reliability* and *Dimensionality* [48]. Both these techniques are used in developing and constructing scales, but are also used to evaluate data that results from using scales in order to confirm the gathered data's validity. A scale's *Dimensionality* measures the number of underlying factors it measures

Demographic	Survey Participants
Age	
18–34	69.2%
35–54	28.9%
>55	1.9%
Gender	
Male	62.3%
Female	37.1%
Other	0.6%
Education	
Some College	45.4%
Bachelor's	45.6%
Master's	8.4%
Other	0.6%

Table 1: Demographic Information of the Participants.

(one, for uni-dimensional, or many, for multi-dimensional scales) and the nature of these factors (correlated or uncorrelated). In our analysis, for each psychometric scale described in Section 3, we conducted a confirmatory factor analysis using Principal Component Analysis (PCA) [49]—applying the rotation method used originally by the authors' of the psychometric scales—to confirm the theoretical structure and underlying factors of the scales. A scale's *Reliability* measures how consistent the scale is when administered repeatedly. We measured reliability using Cronbach's alpha (α) for which, a value of 0.7 of greater is considered acceptable, while a value greater than 0.8 is considered good [50].

Starting with the DoSpeRT scale, a PCA with oblimin rotation revealed the original factor structure of the scale and when taken together, the factors explained 56% of the variance in the data. We extracted these factors and computed their reliability: Ethical ($\alpha = 0.71$), Health/Safety ($\alpha = 0.72$), Financial/Investment ($\alpha = 0.79$), Financial/Gambling ($\alpha = 0.90$), Recreational ($\alpha = 0.82$), and Social ($\alpha = 0.78$).

Next, with the RTC scale, a PCA with oblique rotation revealed the original factor structure of the scale and when taken together, the factors explained 63% of the variance in the data. We extracted these factors and computed their reliability: Emotional Reaction ($\alpha = 0.88$), Short-term Focus ($\alpha = 0.84$), Routine Seeking ($\alpha = 0.81$), and Cognitive Rigidity ($\alpha = 0.70$). Next, we examined the uni-dimensional CFC and the NFC scales. Both these scales revealed high reliability: ($\alpha = 0.9$) and ($\alpha = 0.95$) respectively.

Finally, with SeBIS, a PCA with varimax rotation revealed the original factor structure of the scale and when taken together, the factors explained 59% of the variance in the data. We extracted these factors and computed their reliability: Proactive Awareness ($\alpha = 0.72$), Password Generation ($\alpha = 0.78$), Device Securement ($\alpha = 0.80$), and Device Updating ($\alpha = 0.70$). Therefore, we found sufficient evidence to prove that our data was both reliable and valid.

4.5.2 Logistic Regression: Analyzing Auto-updating Differences

To answer our first research question—investigating the differences that exist between those Android users who currently avoid auto-updating and those who auto-updates their applications—we constructed a logistic regression model [51] using the glm() from the "stats" package in R. Logistic regression is a regression model where the dependent variable is categorical and binary. In our case, the binary outcome is whether participants avoided auto-updating or did auto-update their applications. Because of the low number of participants who restricted auto-updating only for some applications (10%; Section 4.4), we only considered participants' settings from the Play Store (Figure 1) as they reported in Section Two of the survey. We regressed this update choice ("Avoided Auto-updating" = 1 vs "Auto-updated" = 0) on the psychometrics (DoSpeRT, RTC, CFC, NFC), the covariates (SeBIS, age, gender, education), and the presence of a previous negative experience with updating software (Neg. Experience). We coded all those who answered "I don't remember" to having had a previous software updating negative experience as the absence of one. To reduce the number of levels in education, we transformed it into a continuous variable using the following scheme: No High School:1, (High School Graduate, Some College):2, (Bachelor's Degree, Associate's Degree):3, (Master's Degree, Doctoral Degree, Professional Degree):4.

4.5.3 Linear Mixed Effects Model: Analyzing Auto-updating Preferences

To answer our second research question—analyzing how users characteristics explain users' preferences towards auto-updating their mobile applications—we first compiled, for each participant, two scores indicating *how comfortable* they were, one each for security and non-security updates (Update Type), across the data for each application they rated. This lead to a total of $N = 7952$ pairs of (participant, application) with responses to each of the factors we considered in Section 4.1.2. Overall, participants answered the questions about their update preferences for 8.5 applications on average (median = 10). Following that, we constructed a linear mixed-effects regression model [52] using the lmer() from the "lme4" package in R. These models are extensions of the linear regression model in which the predictors contain random effects in addition to the usual fixed effects. Since each participant only answered questions about a subset of the applications—a maximum of 10—we considered a partially crossed random effects model, where the dependent variable was the comfort score, and amongst the predictors, the various user characteristics (Past software updating experience, Psychometric data, SeBIS scores, UpdateType, Trust, Use Frequency, Importance, Satisfied) were the fixed effects, and the subjects and applications were random effects. We also added the demographic variables: age, gender and education, which was coded as before.

5. FINDINGS

Overall, our study showed that users who avoid auto-updates for applications on Android differ from those who auto-update by three characteristics. First, these users have usually had a previous negative experience updating their software—confirming the findings of previous studies for desktop users [12, 13]. However, our study newly suggests that these negative experiences may have occurred on devices other than participants' Android phones, such as their desktops and laptops. Second, these users tend to take fewer financial investment risks (e.g., lower chance to invest money in business ventures and mutual funds) and fewer ethical risks (e.g., lower chance to take questionable deductions on their income task returns). Third, these users exhibit a greater propensity to take proactive steps to maintain their online security—similar to what others have found for desktop users [13]. Overall, across their applications, our participants were less comfortable auto-updating if they had a previous negative updating experience, but more comfortable auto-updating security updates over non-security updates, and applications they deemed trustworthy.

5.1 Differentiating Users Who Avoid Auto updating From Those Who Auto-Update

In our first research question, we asked how those users who avoid auto-updates differ from those who auto-update their applications on Android based on their current reported auto-update settings. The result of the logistic regression model regressing users' update choice on the various psychometric scales, users' past experiences with software updating, and their demographics is shown in Table 2. A likelihood ratio test [53] between the null model and the model with all the predictors revealed an effect size of 0.089 ($p < 0.0001$), and the Nagelkerke R^2 [54] of the model was 0.15—both indicating a good fit for the model.

5.1.1 Past Negative Software Updating Experience

In the survey, nearly 40% of the participants reported having had a previous negative experience with updating their software across their devices. Broken down by whether participants auto-updated their applications, 34.9% of those who auto-updated and 56.8% of those who avoided auto-updating their applications had a previous negative experience with updating their software.

Our results indicate support for hypothesis **H1**, that avoiding auto-updates is associated with past negative experiences with software updates. As seen in Table 2, we observed a significant and large effect size for the coefficient of Negative Experience ($e^\beta = 2.81$, $p < 0.0001$, C.I. = [1.75, 4.56]). Given the positive coefficient, we conclude that having had a previous negative experience with software updating is associated with avoiding auto-updates for applications on Android. It is important to note here that we asked for participants' negative experiences across their devices—not just their Android phones—and this may indicate a tendency for these experiences to affect updating behaviors on devices other than the one they had the negative experience on.

The participants reported a variety of negative experiences with software updating—similar to the reasons reported in previous studies [29, 12, 13, 25]—such as updates caused the software to be buggy, updates made the user interface of the software uncomfortable to use, and updates took a long time to install. Table 3 lists a summary of the negative experiences reported by the survey participants along with the frequency of their appearance. Those participants who chose to elaborate on their negative updating experiences (4.6%) stated:

- How their computers crashed: e.g., P34 *"Windows 10, or garbage time, breaks pretty much every time it updates."*
- How their devices were incompatible with the update: e.g., P145 *"The update I downloaded made other apps buggy."*
- How the update changed and toggled their application settings: e.g., P298 *"The update deleted my password and I could not get it back and it would not let me know what it was. This happened with iTunes and is why I now have Android and not Apple products of any kind. I also lost all the music I had purchased."*

5.1.2 Risk Averse Behavior Intentions

Our results indicated support for hypothesis **H2**, that avoiding auto-updates for applications is associated with lower risk taking behavior, specifically in two domains. First, we found a significant and medium effect size indicating that those who avoid auto-updating their applications also tended to take fewer financial investment risks as indicated by the DoSpeRT-Investment ($e^\beta = 0.79$, $p < 0.01$, C.I. = [0.66, 0.94]) coefficient in Table 2. This means that these users were less likely to invest any money into mutual funds, new business ventures, or speculative stocks. Second, we found a significant and medium effect size indicating that those who avoid auto-updating their applications also tended to take fewer ethical risks as indicated by the DoSpeRT-Ethical ($e^\beta = 0.75$, $p < 0.01$, C.I. = [0.62, 0.91]) coefficient in Table 2. This means that these users were less likely to indulge in affairs with married men/women or keep a lost wallet containing money for themselves. We found no differences in the remaining DoSpeRT domains (Health/Safety, Gambling, Recreational, Social), suggesting that these factors might not differentiate those users who avoid auto-updating from those who auto-update their applications on Android.

At this point, it is worth considering why we found differences particularly in the investment and ethical domains and not the others. We conjecture that because low scores on both the DoSpeRT-Ethical and Investment sub-scales indicate taking responsibility and being in control, they differentiate those users who avoid auto-updating—users with lower scores on these scales—as these users may also express a sense of responsibility over controlling the changes updates make to their devices. These differences may have been less apparent in the DoSpeRT-Gambling sub-scale as gambling and betting are more generally considered risky activities. Furthermore, we conjecture that while both the DoSpeRT-Recreational and Social sub-scales indicate a similar sense of taking responsibility, the associated risks involve references to social activities—that is, interactions with others and in groups—and may therefore, have been less apparent in decisions that affects only the self. Other research [17] has shown that both these sub-scales may be more predictive of users' privacy expectations.

It is also worth pointing out a subtle difference between our result and the results of Egelman and Peer [16] with regards to software updating and risk taking. In their study, Egelman and Peer observed that risk taking was inversely correlated with how often users took actions to update their software (as measured by the SeBIS-Device Updating sub-scale), i.e., low risk taking individuals were more likely to update their software. On the other hand, our results demonstrate that low risk taking individuals were less likely to auto-update—despite the fact that auto-updating should, at least in theory, keep software updated sooner and faster. We believe that this difference is associated with the underlying risks. That is, when users are asked to take actions and make decisions about updating their software, low risk taking individuals are likely to update often because these users are concerned about the risks of not updating (i.e., potential exploits and

Predictor	Estimate	Std. Error	Odds Ratio	Odds Ratio 95% C.I.	p-value
(Intercept)	0.19	1.35	1.21	[0.09, 17.20]	0.89
CFC	−0.33	0.21	0.72	[0.48, 1.09]	0.12
NFC	0.02	0.16	1.02	[0.75, 1.39]	0.92
DoSpeRT-Ethical	−0.29	0.10	0.75	[0.62, 0.91]	**< 0.01**
DoSpeRT-Social	−0.10	0.12	0.90	[0.72, 1.14]	0.39
DoSpeRT-Health/Safety	−0.18	0.14	0.84	[0.64, 1.09]	0.19
DoSpeRT-Recreational	0.17	0.11	1.19	[0.96, 1.48]	0.11
DoSpeRT-Investment	−0.24	0.09	0.79	[0.66, 0.94]	**< 0.01**
DoSpeRT-Gambling	0.08	0.11	1.08	[0.86, 1.34]	0.48
RTC-Emotional Reaction	−0.04	0.18	0.96	[0.69, 1.33]	0.79
RTC-Routine Seeking	−0.03	0.18	0.97	[0.68, 1.38]	0.85
RTC-Cognitive Rigidity	−0.01	0.14	0.99	[0.75, 1.32]	0.95
RTC-Short term Focus	0.17	0.18	1.19	[0.83, 1.70]	0.35
Neg. Experience [Yes]	1.03	0.24	2.81	[1.75, 4.56]	**< 0.0001**
SeBIS-Proactive Awareness	0.35	0.17	1.42	[1.01, 2.01]	**0.04**
SeBIS-Password Generation	0.08	0.16	1.08	[0.79, 1.48]	0.63
SeBIS-Device Updating	−0.13	0.16	0.88	[0.65, 1.19]	0.41
SeBIS-Device Securement	0.10	0.11	1.11	[0.89, 1.39]	0.37
Age	−0.01	0.02	0.99	[0.96, 1.01]	0.33
Education	−0.12	0.19	0.89	[0.61, 1.29]	0.54
Gender [Male]	0.08	0.26	1.09	[0.65, 1.83]	0.74
Model Fit Likelihood Ratio Test: Deviance = 45.99, $p < 0.0001$					
Model Likelihood Ratio Effect Size: 0.089					

Table 2: Results of the Logistic Regression Modeling the Outcome ("Avoided Auto-updating") on the Various Predictors. 95% C.I. is the 95% Confidence Interval. Bolded p-values are Significant at the 0.05 Level.

Negative Experience	Frequency
Version prior to update worked better	36.4%
The update introduced new bugs	34.3%
The update modified the user interface	27.6%
The update took a long time to install	11.3%
The update used up a lot of data	10.7%

Table 3: Negative Experiences with Software Updating Reported by the Survey Participants.

harm). However, with respect to auto-updating, low risk taking individuals are likely to turn off automatic updates because these users are concerned about the risks of auto-updating (i.e., undesirable and negative consequences).

5.1.3 Greater Proactive Security Awareness

Our results also indicated a significant and small effect size indicating that users who avoided auto-updating their applications also exhibited a greater propensity to take proactive steps to maintain their online security; the coefficient of SeBIS-Proactive Awareness ($\beta = 1.42$, $p = 0.04$, C.I. = [1.01, 2.01]). This means that these users were more likely to verify links before opening them, ensure the green HTTPS lock was visible before submitting information, and fix security problems by themselves rather than depending on others. This result supports the findings of Forget *et al.* [13] who, in a study with 15 desktop users, observed that those users who desired control and assumed responsibility in maintaining the security of their computers took proactive security steps (e.g. periodic virus checks), and also sometimes turned off automatic updates.

On the other hand, we observed no such differences between those users who avoided auto-updates and those users who auto-updated their applications with regards to their intentions to behave securely based on the Password Generation, Device Updating, or Device Securement sub-scales—suggesting that the security intentions of users who avoid auto-updates and do auto-update may not be different.

Overall, in our regression model we found evidence to support hypothesis **H1**, auto-updating is associated with previous negative experiences, and hypothesis **H2** that auto-updating applications on Android is associated with taking fewer risks. However, we found no evidence to support hypothesis **H3**, **H4**, **H5**, suggesting that auto-updating may not be associated with consideration of future consequences, curiosity and inquisitiveness, or resistance to change respectively.

5.2 Users' Auto-Updating Preferences

In our second research question, we asked how user characteristics explain users' preferences for how they would like to auto-update across their applications on Android. In the survey, participants rated *how comfortable* they were auto-updating security and non-security updates for the applications they selected. The result of the linear mixed effect regression is shown in Table 4; we only report the significant fixed effects. The $R^2_{LMM_m}$ measure [55] considering only the fixed effects in the model was 0.25—indicating a medium to large effect size. As well, the variance ($\sigma^2 = 432.42$) due to the Participant random effect was much greater than the variance ($\sigma^2 = 5.40$) due to the Application random effect, indicating that differences in preferences towards auto-updating were much greater across participants but fairly consistent across applications.

Fixed Effects	Estimate	Std. Error	Est 95% C.I.	p-value
Neg. Experience [Yes]	-7.39	2.14	$[-11.49, -3.29]$	< 0.001
SeBIS-Proactive Awareness	-3.84	1.47	$[-6.67, -1.02]$	< 0.01
UpdateType [Security]	6.76	0.37	$[6.03, 7.49]$	< 0.0001
Trust	7.29	0.34	$[6.61, 7.96]$	< 0.0001
Importance	2.24	0.25	$[1.76, 2.73]$	< 0.0001
Satisfied	2.96	0.32	$[2.32, 3.58]$	< 0.0001
Model $R^2_{LMM_m}$ Measure: 0.25				

Table 4: Results of the Linear Mixed Effect Model for the Auto-updating Preferences. Participant and Application were Included as Random Effects. 95% C.I. is the 95% Confidence Interval. Only the Significant Fixed Effects are Shown.

Previous Negative Updating Experience: We observed that having had a previous negative experience with software updating affected how comfortable users were towards auto-updating their applications. In our model, we observed a significant and medium effect size for the coefficient of Negative Experience ($\beta = -7.39$, $C.I. = [-11.49, -3.29], p < 0.001$), indicating that once users have a negative experience with updating their software, they become less comfortable auto-updating their applications.

Perceived Trust in the Application: Android applications' perceived trustworthiness played an important part in users' decision making towards auto-updating them. In our model, we observed a significant and medium effect size for the coefficient of Trust ($\beta = 7.29$, $C.I. = [6.61, 7.96], p < 0.0001$), indicating that the more trustworthy users considered an application, the more comfortable they were auto-updating it.

Security Updates vs Non-security Updates: We also observed that the type of update played an important part in whether users would let it apply automatically. In our model, we observed a significant and medium effect size for the coefficient of security updates ($\beta = 6.76$, $C.I. = [6.03, 7.49], p < 0.0001$), indicating that users were more comfortable auto-updating security updates over non-security updates.

Greater Security Awareness: We observed that those users who displayed a higher proactive awareness towards managing their security were less comfortable towards auto-updating their applications. In our model, we observed a significant but small effect size for the coefficient of SeBIS-Proactive Awareness ($\beta = -3.84$, $C.I. = [-6.67, -1.02], p < 0.01$), indicating that users who exhibited greater propensity to engage in proactive security behavior were less comfortable auto-updating their applications.

Perceived Satisfaction with the Application: Android applications' perceived satisfactory performance played a less important part in users' decision making towards auto-updating. In our model, we observed a significant but small effect size for the coefficient of Satisfied ($\beta = 2.96$, $C.I. = [2.32, 3.58], p < 0.0001$), indicating that the more satisfied users were with an application, the more comfortable they were auto-updating it.

Perceived Importance of the Application: Android applications' perceived importance to users also played an important part in users' decision making towards auto-updating them. In our model, we observed a significant but small effect size for the the coefficient of Importance ($\beta = 2.24$, $C.I. = [1.76, 2.73], p < 0.0001$), indicating that the more important users considered an application, the more comfortable they were auto-updating it.

While drawing comparisons with the results for desktop users in [25] is difficult since the previous study included a small sample size, we point out how our results differ. Our results suggest that like desktop users, mobile users are more comfortable auto-updating applications they trust, and more comfortable auto-updating security updates over non-security updates. While desktop users are less comfortable auto-updating applications they are satisfied with and are important to them, both factors have only a small influence on how comfortable mobile users' feel towards auto-updating their applications. Finally, unlike desktop users who were less comfortable auto-updating applications they frequently used, we observed no differences with Android users. It may be possible that we observed no difference because we considered the most popular Android applications which could be used more frequently overall.

6. DISCUSSION

In the following section, we outline the implications of our findings for improving the design of Android OS update system, and encouraging users who avoid auto-updating mobile applications to auto-update security updates.

6.1 Improve Auto-update User Interfaces

To help mobile users keep their applications updated, we suggest that mobile update systems make application software update rollbacks more accessible, and include nudges to encourage users to auto-update security updates.

6.1.1 Make Update Rollbacks Accessible

Our first recommendation stems from our finding that avoiding auto-updating mobile applications on Android is associated with having had a previous negative experience with updating. Therefore, we recommend that one improvement to the current Android OS, or mobile application update systems more generally, would be to provide users with the ability to rollback updates for all applications to a previous point in time to help users who dislike changes made by updates to a particular application to rollback those changes. Presently, updates for applications that are installed from the Play Store cannot be rolled back; update rollbacks are only allowed for applications that come pre-installed with the device (as described in Section 2.3). However, because the implications of update rollbacks may be potentially harmful—as security updates may be rolled back as well—this change may require a mobile OS to restrict update rollback for applications that do not contain recent security updates. Such a system would also rely on informing users about potential feature losses to help them understand what rolling back updates would entail. Overall, this would potentially increase end-users' confidence in auto-updating security updates.

6.1.2 Design Nudges for Auto-Updating

Our second recommendation stems from our findings that users who avoid auto-updating mobile applications on Android also tend to take fewer investment risks and fewer ethical risks. We suggest that these characteristics can be used to design "nudges" to persuade users to auto-update security updates, since our findings suggest that overall, users are more comfortable automating security updates. Nudges entail the use of behavioral economics to encourage users into making certain decisions [56]. Numerous studies have experimented with nudging to affect behavior change in domains ranging from health [57, 58, 59] to retirement savings policies [60, 61]. In the privacy and security community, recent studies have created nudges for users to make stronger passwords using password meters [62], and others have nudged users into reducing regrets during online social network use [63].

Specifically, we envision nudges to encourage users to auto-update security updates that leverage the vast literature on "Framing Effects" [64]—a cognitive bias in which people react to choices based on how these choices are framed (such as loss vs gains). For instance, because users who avoid auto-updating application updates on Android also take fewer ethical risks as measured by the DoSpeRT-Ethical scale, one nudge could highlight the ethical risk and responsibility associated with not auto-updating security updates, e.g., *"Switching auto-updates on for security updates will protect you, and others like you from suffering the consequences of someone exploiting your device"*. Similarly, because the same users also take fewer investment risks, as measured by the DoSpeRT-Investment scale, another nudge could highlight the financial and investment risk associated with not auto-updating security updates: *"Not switching auto-updates on for security updates increases the chances of someone gaining access to your bank account or stealing your credit card information"*. Both these nudges, using the risk taking traits resulting from our findings, emphasize the potential losses—an attacker exploiting your device, or accessing your credit card information—that may result from not auto-updating security updates.

Such nudges could also emphasize that negative experiences such as changes to the user interface or data loss, will be minimized since these are primarily security updates. If the Android OS implements a mechanism to rollback updates, as we touched upon in Section 6.1.1, these nudges could also remind users about the application update rollback setting and emphasize that users can un-install updates at any time if they dislike the changes caused post update. These nudges could be presented to users at different times. For instance, these nudges could be presented to users who avoid auto-updates when they attempt to switch off auto-updates, or soon after they manually install an update for an application. Future research could test the effectiveness of these nudges and messages by means of various controlled experiments.

In addition to nudges, security education experts can leverage the same ethical and investment risk taking trait differences between users who avoid auto-updating and those who do not in order to design better security education campaigns and security advice for end-users. For instance, the ethical risk taking could be used to highlight that *"Users have a responsibility to auto-update their systems and keep their organization and fellow users safe"*, and changes in user behavior could be measured pre- and post-training.

Of course, allowing auto-updates for security updates only will entail providing software developers with the incentives, education, and necessary infrastructure to decouple mobile application security updates, whenever possible, from all other kinds of updates. On Android, enabling this functionality would also require a redesign of the application update interface. As Figure 1 shows, the Android OS currently has provisions for users to either auto-update all their applications, disallow auto-update for certain applications (through the applications' page on the Play Store), or auto-update none of their applications. However, there is no provision for users to automate certain kinds of updates over others. In contrast, the Apple Mac OS X system allows users to selectively automate security OS updates while restricting updates of all other types [65]. The Android OS could extend this concept to the Play Store, providing users with another option to automate only security updates.

6.2 Examine Update Development Practices

Our third recommendation stems again from our finding that mobile users who avoid application auto-updates have had a previous negative experience with software updating. We suggest that the burden of updating applications and device should not solely rely on end-users, and echoing the call of others [66], we hope that the security community go beyond studying the updating behaviors of end-users and also investigate how software developers decide to develop, build, and test software updates in the first place. Like end-users, software developers make trade-offs when deciding what content to add or remove via an update, or what security changes to push to end-users. These trade-offs maybe influenced by a variety of factors, including their attitudes, motivations, and the feedback they receive from their end-users. Future research could identify how software developers propagate application changes to their end-users, what specific changes lead to negative experiences for end-users, and ways to minimize these downsides as part of the update development process. For instance, when developers add or remove features from a particular application, how do they consider these changes will impact those users with the current version of the application, and how do developers decide what information to provide to users to inform them about the changes made by updates?

6.3 Personalize Mobile Auto-update Systems

Our fourth recommendation stems from our finding that overall, mobile users' perceived level of trust with applications, and the type of update (security over non-security updates) is positively and strongly correlated with how comfortable they felt auto-updating their applications. Because users' auto-updating preferences contain some nuances, we argue that a one-size-fits-all update system may be less optimal, and work against the preferences of those users who avoid auto-updating. While the Play Store does provide users with the ability to restrict auto-updates for certain applications—

as described in Section 2.3—the choice to do so lies with the end users, and can ultimately be an effort requiring task because a user can have 95 different applications installed on average [67]. Indeed, in our data only 10% of our participants actually used this feature.

Therefore, we propose that mobile update systems need to be more personalized and learn from users and their actions, and accordingly decide which applications to auto-update and which others to not—selectively involving users only when necessary. Our findings provide a starting point for personalization based on the user characteristics and preferences we identified for users auto-updating their mobile applications. For instance, since users in our study were more comfortable auto-updating applications they trusted, future systems could explore and uncover proxies for trust, and use that to drive auto-update decisions. Some proxies of trust might include dimensions such as whether a user provides a high rating, or a positive comment for a particular application, or has downloaded multiple applications from the same application developer.

These proxies could allow the system to help suggest or even decide when to automatically install updates for any particular application depending on the user's preference for consent to update. As another example, the system could automatically install all updates for applications that users might generally trust, such as emergency applications. Of course, such a system would require great transparency, and be able to inform users about *what* actions it has taken and *how* it arrived at the decision of taking those actions.

7. LIMITATIONS

Our study has several limitations. First, our results correlating user characteristics and their choice of automating updates is limited to how applications are updated on the Android platform, and therefore limited in how far they can be generalized to non-application updates (such as OS updates). Second, as noted in Section 4.3, our survey collected users' self-reported update settings, and these settings may be subject to error. Third, the applications users reported having installed on their phones could be subject to recall bias, and are only limited to the ones we presented as part of the survey. However, providing a list of all applications was impractical, and we had to pare it to the most popular applications. Fourth, as a result of choosing the AMT platform, our results are limited in their generalizability to other Android users. However, while the psychometric scales have been used outside of AMT before, the SeBIS scale has only been tested and validated on AMT, which made it a reasonable platform to run our study. Furthermore, the AMT population, while limited in its diversity, has been shown to be fairly similar to participants from university campuses and other online participant pools [68, 69]. Despite these limitations, our study provides insight into users preferences towards auto-updating mobile applications.

8. CONCLUSION AND FUTURE WORK

We conducted a survey to understand how user characteristics affect attitudes towards mobile application updates on Android. We found that three characteristics differentiated those users who avoid auto-updates from those who auto-update their mobile applications. These characteristics are past experiences with software updating, propensity to engage in risk taking behavior, and displaying greater proactive awareness about their online security.

We also found that previous negative experiences made users less comfortable with auto-updating their applications. However, users were more comfortable with the idea of auto-updating security updates and applications they deemed more trustworthy. Based on these findings, we made four recommendations for improving security on Android by encouraging users to switch on auto-updates via making application update rollbacks more accessible, nudging users to auto-update, studying software developers and their update development practices, and using our findings as a starting point for personalizing mobile update systems.

Future work could examine how users' attitudes towards auto-updating vary on other platforms and devices, and more directly observe or infer how users update as opposed to using self-reported user data. Future work could also examine how our results generalize beyond the AMT platform by repeating our survey on a more representative sample of Android users. Finally, future work could use controlled experiments to present users with different versions of the nudges we proposed, and measure whether or not users are moved to switch on auto-updates after being exposed to these kinds of nudges. Another potential area for future inquiry would be to build on our findings to help create both user and application profiles for personalizing auto-updates in mobile OS update systems.

Lastly, auto-updating applications on a mobile also poses an interesting dichotomy: while on one hand auto-updates may bring enhanced security and protection, on the other hand these updates can also be abused by malicious software developers. These malicious developers might want to use this channel to collect more data about users that might be of high value to advertisers, or inject advertisement libraries. A final suggestion for future research is to consider this dichotomy in greater detail, and devise ways so that software updates are vetted before they can be automated.

9. ACKNOWLEDGEMENTS

We thank Yasemin Acar for shepherding the paper, and the anonymous reviewers for their helpful comments. We also thank Michelle Mazurek, Jessica Vitak, Marian Harbach, Nathan Malkin, and Elissa Redmiles for feedback on earlier drafts of the paper. Our research is based upon work supported by the Maryland Procurement Office under contract H98230-14-C-0137. 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 Maryland Procurement Office.

10. REFERENCES

[1] Symantec. Internet Security Threat Report. `https://www.symantec.com/content/dam/symantec/docs/reports/istr-21-2016-en.pdf`, April 2016.

[2] Kenneth Olmstead and Aaron Smith. Americans and Cybersecurity. `http://www.pewinternet.org/2017/01/26/americans-and-cybersecurity/`, January 2017.

[3] 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, Washington,

D.C., 2015. USENIX Association.

[4] Net Applications Inc. Mobile/Tablet Operating System Market Share. https://www.netmarketshare.com/operating-system-market-share.aspx?qprid=8&qpcustomd=1, August 2016.

[5] Sascha Fahl, Marian Harbach, Thomas Muders, Lars Baumgärtner, Bernd Freisleben, and Matthew Smith. Why Eve and Mallory Love Android: An Analysis of Android SSL (in)Security. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, CCS '12, pages 50–61, New York, NY, USA, 2012. ACM.

[6] Vishwanath Raman Adrian Mettler, Yulong Zhang. SSL Vulnerabilities: Who Listens When Android Applications Talk? https://www.fireeye.com/blog/threat-research/2014/08/ssl-vulnerabilities-who-listens-when-android-\applications-talk.html, August 2014.

[7] Williams Pelegrin. These android, ios, and wp8 apps are affected by the heartbleed bug. http://www.digitaltrends.com/mobile/heartbleed-bug-apps-affected-list/, April 2014.

[8] Cisco. Cisco Annual Security Report. https://www.cisco.com/web/offers/pdfs/cisco-asr-2015.pdf, 2015.

[9] Antonio Nappa, Richard Johnson, Leyla Bilge, Juan Caballero, and Tudor Dumitras. The Attack of the Clones: A Study of the Impact of Shared Code on Vulnerability Patching. In *Security and Privacy (SP), 2015 IEEE Symposium on*, pages 692–708, Piscataway, NJ, USA, May 2015. IEEE.

[10] US-CERT. Before You Connect a New Computer to the Internet. https://www.us-cert.gov/ncas/tips/ST15-003, December 2015.

[11] Thomas Duebendorfer and Stefan Frei. Why silent updates boost security. *TIK, ETH Zurich, Tech. Rep*, 302, 2009.

[12] Kami Vaniea and Yasmeen Rashidi. Tales of Software Updates: The Process of Updating Software. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 3215–3226, New York, NY, USA, 2016. ACM.

[13] 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, Denver, CO, June 2016. USENIX Association.

[14] Rick Wash and Emilee Rader. Too Much Knowledge? Security Beliefs and Protective Behaviors Among United States Internet Users. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 309–325, Ottawa, 2015. USENIX Association.

[15] Monica Whitty, James Doodson, Sadie Creese, and Duncan Hodges. Individual differences in cyber security behaviors: an examination of who is sharing passwords. *Cyberpsychology, Behavior, and Social Networking*, 18(1):3–7, 2015.

[16] 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*, CHI '15, pages 2873–2882, New York, NY, USA, 2015. ACM.

[17] Serge Egelman and Eyal Peer. The Myth of the Average User: Improving Privacy and Security Systems Through Individualization. In *Proceedings of the 2015 New Security Paradigms Workshop*, NSPW '15, pages 16–28, New York, NY, USA, 2015. ACM.

[18] Juan Herrero, Alberto Urueña, Andrea Torres, and Antonio Hidalgo. My computer is infected: the role of users' sensation seeking and domain-specific risk perceptions and risk attitudes on computer harm. *Journal of Risk Research*, pages 1–14, 2016.

[19] M. Tischer, Z. Durumeric, S. Foster, S. Duan, A. Mori, E. Bursztein, and M. Bailey. Users really do plug in usb drives they find. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 306–319, May 2016.

[20] Nathan Malkin, Arunesh Mathur, Marian Harbach, and Serge Egelman. Personalized security messaging: Nudges for compliance with browser warnings. In *2nd European Workshop on Usable Security*. Internet Society, 2017.

[21] Pam Briggs Debora Jeske, Lynne Coventry and Aad van Moorsel. Nudging whom how: IT proficiency, impulse control and secure behaviour. *Networks*, 49:18, 2014.

[22] Vaibhav Garg, L. Jean Camp, Katherine Connelly, and Lesa Lorenzen-Huber. *Risk Communication Design: Video vs. Text*, pages 279–298. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[23] Timothy Kelley, L. Jean Camp, Suzanne Lien, and Douglas Stebila. Self-identified Experts Lost on the Interwebs: The Importance of Treating All Results As Learning Experiences. In *Proceedings of the 2012 Workshop on Learning from Authoritative Security Experiment Results*, LASER '12, pages 47–54, New York, NY, USA, 2012. ACM.

[24] 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 SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 373–382, New York, NY, USA, 2010. ACM.

[25] Arunesh Mathur, Josefine Engel, Sonam Sobti, Victoria Chang, and Marshini Chetty. "They Keep Coming Back Like Zombies": Improving Software Updating Interfaces. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 43–58, Denver, CO, June 2016. USENIX Association.

[26] 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*, CHI '16, pages 5257–5261, New York, NY, USA, 2016. ACM.

[27] Iulia Ion, Rob Reeder, and Sunny Consolvo. "...No one Can Hack My Mind": Comparing Expert and Non-Expert Security Practices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages

327–346, Ottawa, July 2015. USENIX Association.

[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 *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pages 666–677, New York, NY, USA, 2016. ACM.

[29] Kami E. Vaniea, Emilee Rader, and Rick Wash. Betrayed by Updates: How Negative Experiences Affect Future Security. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 2671–2674, New York, NY, USA, 2014. ACM.

[30] Michael Fagan, Mohammad Maifi Hasan Khan, and Ross Buck. A Study of Users' Experiences and Beliefs about Software Update Messages. *Computers in Human Behavior*, 51, Part A:504 – 519, 2015.

[31] Marshini Chetty, Richard Banks, A.J. Brush, Jonathan Donner, and Rebecca Grinter. You're Capped: Understanding the Effects of Bandwidth Caps on Broadband Use in the Home. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 3021–3030, New York, NY, USA, 2012. ACM.

[32] Arunesh Mathur, Brent Schlotfeldt, and Marshini Chetty. A Mixed-methods Study of Mobile Users' Data Usage Practices in South Africa. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 1209–1220, New York, NY, USA, 2015. ACM.

[33] Andreas Moller, Stefan Diewald, Luis Roalter, Technische Universitat Muchen, Florian Michahelles, and Matthias Kranz. Update Behavior in App Markets and Security Implications: A Case Study in Google Play. In *In Proc. of the 3rd Intl. Workshop on Research in the Large. Held in Conjunction with Mobile HCI*, pages 3–6, 2012.

[34] Yuan Tian, Bin Liu, Weisi Dai, Blase Ur, Patrick Tague, and Lorrie Faith Cranor. Supporting privacy-conscious app update decisions with user reviews. In *Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices*, SPSM '15, pages 51–61, New York, NY, USA, 2015. ACM.

[35] Christos Gkantsidis, Thomas Karagiannis, and Milan VojnoviC. Planet Scale Software Updates. *SIGCOMM Comput. Commun. Rev.*, 36(4):423–434, August 2006.

[36] Rick Wash, Emilee Rader, Kami Vaniea, and Michelle Rizor. Out of the Loop: How Automated Software Updates Cause Unintended Security Consequences. In *Tenth Symposium On Usable Privacy and Security (SOUPS 2014)*, pages 89–104, Menlo Park, 2014. USENIX Association.

[37] W. Keith Edwards, Erika Shehan Poole, and Jennifer Stoll. Security Automation Considered Harmful? In *Proceedings of the 2007 Workshop on New Security Paradigms*, NSPW '07, pages 33–42, New York, NY, USA, 2008. ACM.

[38] Google. Update Downloaded Apps. `https://support.google.com/googleplay/answer/113412?hl=en`, September 2016.

[39] Rollback or Uninstall Updates on Android App. `https://www.updateallapps.com/rollback-uninstall-updates-android-app/`, June 2016.

[40] Google. Working with System Permissions. `https://developer.android.com/training/permissions/index.html`, 2016.

[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] Ann-Renée Blais and Elke U Weber. A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, 1(1), 2006.

[43] Jeff Joireman, Monte J Shaffer, Daniel Balliet, and Alan Strathman. Promotion orientation explains why future-oriented people exercise and eat healthy evidence from the two-factor consideration of future consequences-14 scale. *Personality and Social Psychology Bulletin*, 38(10):1272–1287, 2012.

[44] Richard E Petty, John T Cacioppo, and Chuan Feng Kao. The efficient assessment of need for cognition. *Journal of Personality Assessment*, 48(3):306–307, 1984.

[45] Shaul Oreg. Resistance to change: developing an individual differences measure. *Journal of applied psychology*, 88(4):680, 2003.

[46] Wikipedia. List of Most Downloaded Android Applications. `https://en.wikipedia.org/wiki/List_of_most_downloaded_Android_applications`, 2016.

[47] Seymour Sudman, Norman M Bradburn, and Norbert Schwarz. *Thinking about answers: The application of cognitive processes to survey methodology.* Jossey-Bass, 1996.

[48] Mike Furr. *Scale construction and psychometrics for social and personality psychology.* SAGE Publications Ltd, 2011.

[49] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

[50] Rosemary R Gliem and Joseph A 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.

[51] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[52] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*, 2014.

[53] Scott Menard. Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54(1):17–24, 2000.

[54] Nico JD Nagelkerke. A note on a general definition of the coefficient of determination. *Biometrika*, 78(3):691–692, 1991.

[55] Shinichi Nakagawa and Holger Schielzeth. A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142, 2013.

[56] H Thaler Richard and R Sunstein Cass. Nudge: Improving decisions about health, wealth, and happiness, 2008.

[57] Eric J Johnson and Daniel G Goldstein. Defaults and donation decisions. *Transplantation*, 78(12):1713–1716, 2004.

[58] Julie S. Downs, George Loewenstein, and Jessica Wisdom. Strategies for Promoting Healthier Food Choices. *American Economic Review*, 99(2):159–64, May 2009.

[59] Scott D. Halpern, Peter A. Ubel, and David A. Asch. Harnessing the Power of Default Options to Improve Health Care. *New England Journal of Medicine*, 357(13):1340–1344, 2007. PMID: 17898105.

[60] Brigitte C Madrian and Dennis F Shea. The power of suggestion: Inertia in 401 (k) participation and savings behavior. *The Quarterly Journal of Economics*, 116(4):1149–1187, 2001.

[61] James J Choi, David Laibson, Brigitte C Madrian, Andrew Metrick, et al. Saving for retirement on the path of least resistance. *Rodney L White Center For Financial Research - Working Papers* -, 9, 2005.

[62] Serge Egelman, Andreas Sotirakopoulos, Ildar Muslukhov, Konstantin Beznosov, and Cormac Herley. Does My Password Go Up to Eleven?: The Impact of Password Meters on Password Selection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2379–2388, New York, NY, USA, 2013. ACM.

[63] Yang Wang, Pedro Giovanni Leon, Alessandro Acquisti, Lorrie Faith Cranor, Alain Forget, and Norman Sadeh. A Field Trial of Privacy Nudges for Facebook. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 2367–2376, New York, NY, USA, 2014. ACM.

[64] Irwin P Levin, Sandra L Schneider, and Gary J Gaeth. All frames are not created equal: A typology and critical analysis of framing effects. *Organizational behavior and human decision processes*, 76(2):149–188, 1998.

[65] Apple. Mac App Store: Automatic security updates. `https://support.apple.com/en-us/HT204536`, September 2016.

[66] 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. IEEE, 2016.

[67] Paul Sawers. Android users have an average of 95 apps installed on their phones, according to yahoo aviate data. `https://thenextweb.com/apps/2014/08/26/android-users-average-95-apps-installed\-phones-according-yahoo-aviate-data/`, August 2014.

[68] Daniel J Simons and Christopher F Chabris. Common (mis) beliefs about memory: A replication and comparison of telephone and mechanical turk survey methods. *PloS one*, 7(12):e51876, 2012.

[69] Christoph Bartneck, Andreas Duenser, Elena Moltchanova, and Karolina Zawieska. Comparing the similarity of responses received from studies in amazon's mechanical turk to studies conducted online and with direct recruitment. *PloS one*, 10(4):e0121595, 2015.

APPENDIX

A. PART ONE: PSYCHOMETRIC SCALES

1. Domain Specific Risk Taking (DoSpeRT) scale [42] (Ethical = E, Financial/Investment = F/I, Financial / Gambling = F/G, Health/Safety = HS, Social = S, Recreational = R) [Scoring: 1 (Extremely Unlikely) and 7 (Extremely Likely)]
   - Admitting that your tastes are different from those of a friend. (S)
   - Disagreeing with an authority figure on a major issue. (S)
   - Choosing a career that you truly enjoy over a more secure one. (S)
   - Speaking your mind about an unpopular issue in a meeting at work. (S)
   - Moving to a city far away from your extended family. (S)
   - Starting a new career in your mid-thirties. (S)
   - Going camping in the wilderness. (R)
   - Taking a skydiving class. (R)
   - Bungee jumping off a tall bridge. (R)
   - Piloting a small plane. (R)
   - Going down a ski run that is beyond your ability. (R)
   - Going whitewater rafting at high water in the spring. (R)
   - Betting a day's income at the horse races. (F/G)
   - Betting a day's income at a high-stake poker game. (F/G)
   - Betting a day's income on the outcome of a sporting event. (F/G)
   - Investing 10% of your annual income in a moderate growth mutual fund. (F/I)
   - Investing 5% of your annual income in a very speculative stock. (F/I)
   - Investing 10% of your annual income in a new business venture. (F/I)
   - Drinking heavily at a social function. (HS)
   - Sunbathing without sunscreen. (HS)
   - Riding a motorcycle without a helmet. (HS)
   - Driving or riding a car without wearing a seat belt. (HS)
   - Walking home alone at night in an unsafe area of town. (HS)
   - Engaging in unprotected sex. (HS)
   - Taking some questionable deductions on your income tax return. (E)
   - Having an affair with a married man/woman. (E)
   - Passing off somebody else's work as your own. (E)
   - Revealing a friend's secret to someone else. (E)
   - Not returning a wallet you found that contains $200. (E)
   - Leaving your young children alone at home while running an errand. (E)

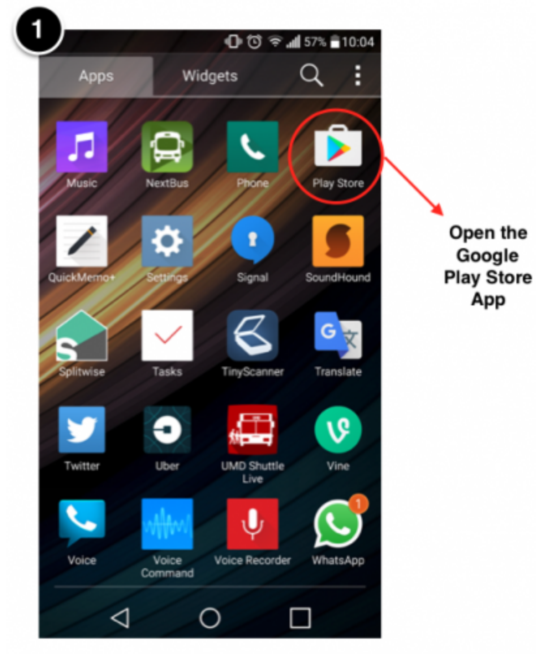
2. Consideration for Future Consequences (CFC) scale [43] [Scoring: 1 (Extremely Uncharacteristic of Me) and 7 (Extremely Characteristic of Me)]
   - I consider how things might be in the future, and try to influence those things with my day to day behavior.
   - Often I engage in a particular behavior in order to achieve outcomes that may not result for many years.
   - I only act to satisfy immediate concerns, figuring the future will take care of itself.
   - My behavior is only influenced by the immediate (i.e., a matter of days or weeks) outcomes of my actions.
   - My convenience is a big factor in the decisions I make or the actions I take.
   - I am willing to sacrifice my immediate happiness or well-being in order to achieve future outcomes.
   - I think it is important to take warnings about negative outcomes seriously even if the negative outcome will not occur for many years.
   - I think it is more important to perform a behavior with important distant consequences than a behavior with less important immediate consequences.
   - I generally ignore warnings about possible future problems because I think the problems will be resolved before they reach crisis level.
   - I think that sacrificing now is usually unnecessary since future outcomes can be dealt with at a later time.
   - I only act to satisfy immediate concerns, figuring that I will take care of future problems that may occur at a later date.
   - Since my day-to-day work has specific outcomes, it is more important to me than behavior that has distant outcomes.
   - When I make a decision, I think about how it might affect me in the future.
   - My behavior is generally influenced by future consequences.
3. Need for Cognition (NFC) scale [44] [Scoring: 1 (Extremely Uncharacteristic of Me) and 5 (Extremely Characteristic of Me)]
   - I would prefer complex to simple problems.
   - I like to have the responsibility of handling a situation that requires a lot of thinking.
   - Thinking is not my idea of fun.
   - I would rather do something that requires little thought than something that is sure to challenge my thinking abilities.
   - I try to anticipate and avoid situations where there is a likely chance I will have to think in depth about something.
   - I find satisfaction in deliberating hard and for long hours.
   - I only think as hard as I have to.
   - I prefer to think about small daily projects to long term ones.
   - I like tasks that require little thought once I've learned them.
   - The idea of relying on thought to make my way to the top appeals to me.
   - I really enjoy a task that involves coming up with new solutions to problems.
   - Learning new ways to think doesn't excite me very much.
   - I prefer my life to be filled with puzzles I must solve.
   - The notion of thinking abstractly is appealing to me.
   - I would prefer a task that is intellectual, difficult, and important to one that is somewhat important but does not require much thought.
   - I feel relief rather than satisfaction after completing a task that requires a lot of mental effort.
   - It's enough for me that something gets the job done; I don't care how or why it works.
   - I usually end up deliberating about issues even when they do not affect me personally.
4. Resistance to Change (RTC) scale [45] (RS = Routine Seeking, ER = Emotional Reaction, SF = Short-term Focus, CR = Cognitive Rigidity) [Scoring: 1 (Strongly Disagree) and 6 (Strongly Agree)]
   - I generally consider changes to be a negative thing. (RS)
   - I'll take a routine day over a day full of unexpected events any time. (RS)
   - I like to do the same old things rather than try new and different ones. (RS)
   - Whenever my life forms a stable routine, I look for ways to change it. (RS)
   - I'd rather be bored than surprised. (RS)
   - If I were to be informed that there's going to be a significant change regarding the way things are done at work, I would probably feel stressed. (ER)
   - When I am informed of a change of plans, I tense up a bit. (ER)
   - When things don't go according to plans, it stresses me out. (ER)
   - If one of my bosses changed the performance evaluation criteria, it would probably make me feel uncomfortable even if I thought I'd do just as well without having to do any extra work. (ER)
   - Changing plans seems like a real hassle to me. (SF)
   - Often, I feel a bit uncomfortable even about changes that may potentially improve my life. (SF)
   - When someone pressures me to change something, I tend to resist it even if I think the change may ultimately benefit me. (SF)
   - I sometimes find myself avoiding changes that I know will be good for me. (SF)
   - I often change my mind. (CR)
   - I don't change my mind easily. (CR)
   - Once I've come to a conclusion, I'm not likely to change my mind. (CR)
   - My views are very consistent over time. (CR)

5. Security Behavior Intentions (SeBIS) scale [16] (DU = Device Updating, DS = Device Securement, PG = Password Generation, PA = Proactive Awareness) [Scoring: 1 (Never) and 5 (Always)]

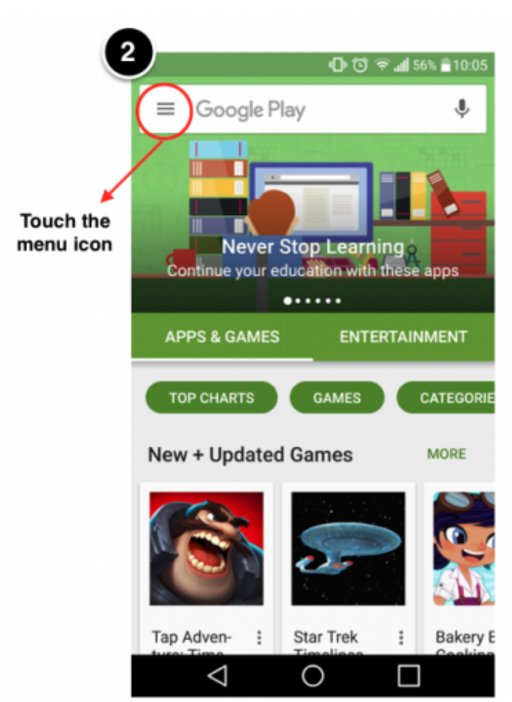
- I set my computer screen to automatically lock if I don't use it for a prolonged period of time. (DS)
- I use a password/passcode to unlock my laptop or tablet. (DS)
- I manually lock my computer screen when I step away from it. (DS)
- I use a PIN or passcode to unlock my mobile phone. (DS)
- I do not change my passwords, unless I have to. (PG)
- I use different passwords for different accounts that I have. (PG)
- When I create a new online account, I try to use a password that goes beyond the site's minimum requirements. (PG)
- I do not include special characters in my password if it's not required. (PG)
- When someone sends me a link, I open it without first verifying where it goes. (PA)
- I know what website I'm visiting based on its look and feel, rather than by looking at the URL bar. (PA)
- I submit information to websites without first verifying that it will be sent securely (e.g., SSL, "https://", a lock icon). (PA)
- When browsing websites, I mouseover links to see where they go, before clicking them. (PA)
- If I discover a security problem, I continue what I was doing because I assume someone else will fix it. (PA)
- When I'm prompted about a software update, I install it right away. (DU)
- I try to make sure that the programs I use are up-to-date. (DU)
- I verify that my anti-virus software has been regularly updating itself. (DU)

B. PART TWO: ANDROID APPLICATION UPDATE SETTINGS AND AUTO-UPDATING PREFERENCES

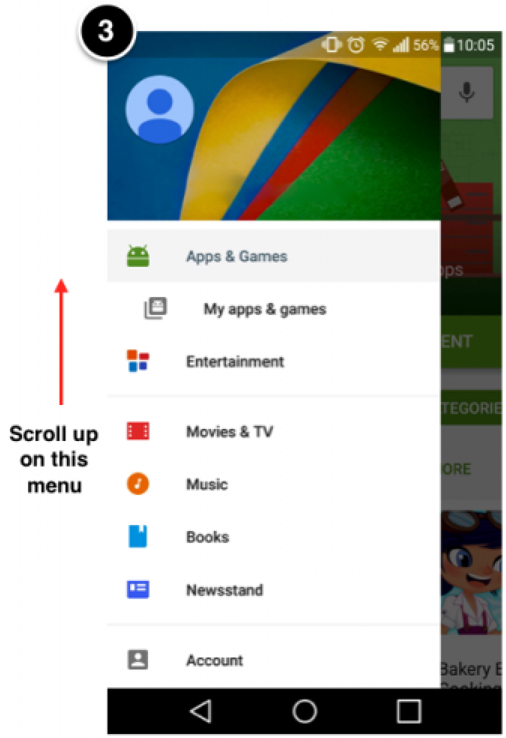
1. Please report the following update settings for your Android device by following the instructions in the images below.



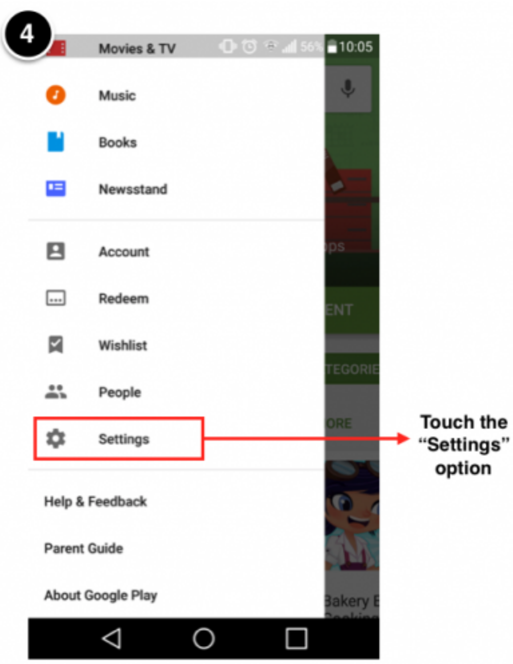
1.



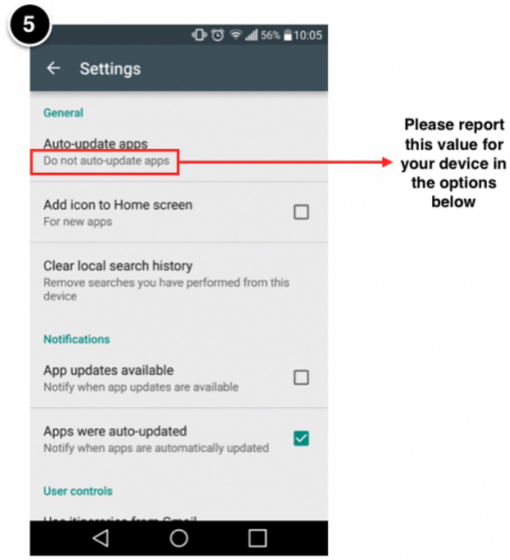
2.



3.



4.

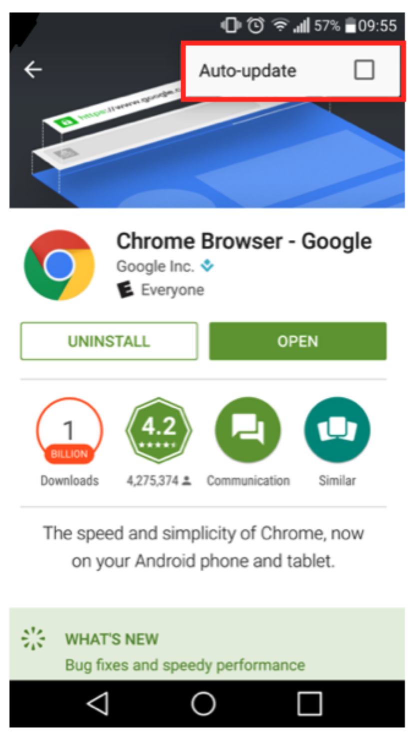


5.

Please report the text in image (5) for your device:

1. Do not auto-update apps.
2. Auto-update apps at any time. Data charges may apply.
3. Auto-update apps over Wi-Fi only.
4. I don't know

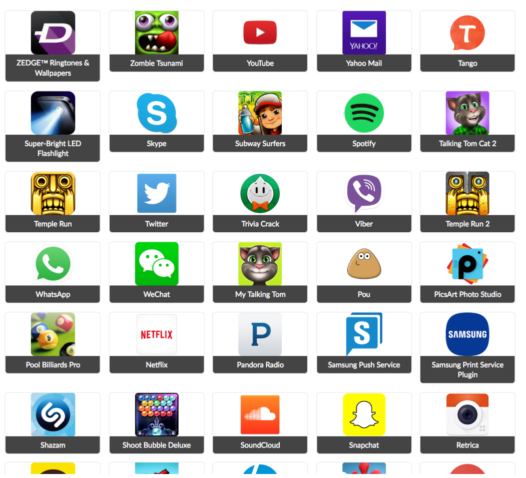
2. The Google Play Store allows certain apps to be updated manually. For example, the following images describe how by deselecting the Auto-update checkbox, the Google Chrome app will no longer be auto-updated. [Only shown if the answer to the previous questions is (b) or (c)]



Do you manually update certain apps in the manner shown above?

1. I do not manually update any of my apps in this manner
2. I manually update some of my apps in this manner
3. I manually update most of my apps in this manner
4. I don't know

3. The following is a list of the most downloaded Android apps from the Google Play Store. From this list, please select ALL the ones you have installed on your Android phone. List taken from [46].



For a maximum of 10 randomly selected applications from the previous question:

1. Assuming no data charges apply, how comfortable are you setting security updates to automatically download and install for the following apps? [0 - 100]
2. Assuming no data charges apply, how comfortable are you setting NON security updates to automatically download and install for the following apps? [0 - 100]
3. How frequently do you use the following apps? [Several times a day (5) - Less often (1)]
4. How trustworthy do you feel are the following apps? [Not at all trustworthy (1) - Extremely trustworthy (5)]
5. How satisfied are you with using the following apps? [Not at all satisfied (1) - Extremely satisfied (5)]
6. How important are the following apps to you? [Not at all important (1) - Extremely important (5)]

C. PART THREE: PAST UPDATE EXPERIENCES

1. Have you ever regretted or had a negative experience updating any software across your devices?
   (a) Yes
   (b) No
   (c) I don't remember
2. The following are some reasons why people regret installing updates. Please check all the reasons that have caused you to regret updating your software. [Only shown if the answer to the previous questions is "Yes"]
   (a) The update introduced new bugs in the software.
   (b) The update changed the user interface.
   (c) The update used up a lot of data.
   (d) The update took more time to install than I expected it to take.
   (e) The old version of the software worked better than the updated one.
   (f) Other - Write In

D. DEMOGRAPHICS

1. What is your age? [Write In]
2. What is your annual household income?
   (a) Less than $25,000
   (b) $25,000 to $34,999
   (c) $35,000 to $49,999
   (d) $50,000 to $74,999
   (e) $75,000 to $99,999
   (f) $100,000 to $124,999
   (g) $125,000 to $149,999
   (h) $150,000 or more
   (i) Prefer not to answer
3. What is the highest education level you have completed?
   (a) No High School
   (b) High School Graduate
   (c) Some College
   (d) Bachelor's Degree
   (e) Associate's Degree
   (f) Master's Degree
   (g) Doctoral Degree
   (h) Professional Degree (e.g., MBA, J.D.)
   (i) Prefer not to answer
4. What gender do you most closely identify with?
   (a) Male
   (b) Female
   (c) Other
   (d) Prefer not to answer

Exploring decision making with Android's runtime permission dialogs using in-context surveys

Bram Bonné
Hasselt University - tUL - imec
bram.bonne@uhasselt.be

Sai Teja Peddinti, Igor Bilogrevic, Nina Taft
Google Inc.
{psaiteja, ibilogrevic, ninataft}@google.com

ABSTRACT

A great deal of research on the management of user data on smartphones via permission systems has revealed significant levels of user discomfort, lack of understanding, and lack of attention. The majority of these studies were conducted on Android devices before runtime permission dialogs were widely deployed. In this paper we explore how users make decisions with runtime dialogs on smartphones with Android 6.0 or higher. We employ an experience sampling methodology in order to ask users the reasons influencing their decisions immediately after they decide. We conducted a longitudinal survey with 157 participants over a 6 week period.

We explore the grant and denial rates of permissions, overall and on a per permission type basis. Overall, our participants accepted 84% of the permission requests. We observe differences in the denial rates across permissions types; these vary from 23% (for microphone) to 10% (calendar). We find that one of the main reasons for granting or denying a permission request depends on users' expectation on whether or not an app should need a permission. A common reason for denying permissions is because users know they can change them later. Among the permissions granted, our participants said they were comfortable with 90% of those decisions - indicating that for 10% of grant decisions users may be consenting reluctantly. Interestingly, we found that women deny permissions twice as often as men.

1. INTRODUCTION

Mobile users have an immense choice when searching for an app to install on their devices. Two of the most popular mobile platforms, Google's Android and Apple's iOS, each have more than a million different third-party apps that users can choose from [45], not to mention the additional third-party marketplaces. Users make a number of decisions during the lifecycle of an app on their smartphones, including deciding to install an app, making choices about whether or not to give an app access to personal data, and potentially uninstalling the app. There are many factors that could commingle to bring users to a decision. Part of the thinking around these decisions may involve reasons related to privacy, such as sensitivity to sharing particular types of data, trust in the developer, understanding the value added when personal data is shared, and many more [18, 19, 20, 29]. In order for an app to access personal data, both Android and iOS adopt a runtime permission model, which allows users to decide whether to grant a given permission request at the time when it is first needed within the app. In this paper we explore users' rationales for decision making during these three parts of an app's lifecycle, but with a focus on how users decide about permissions. Importantly, we study users' rationales at the moment they make their decision.

A large body of work has focused on understanding users' attitudes, comfort and their comprehension about permissions [2, 4, 15, 24]. However, almost all prior studies were conducted by using the permission model in which users had to accept or deny all the permissions requested by an app at installation time, without the possibility to grant permissions individually (for versions of Android before 6.0). A series of notable findings by Felt et al. [15] and Kelley et al. [24] showed that few users pay attention to install-time permission dialogs and even fewer understand them. Furthermore, results from other studies [2, 4, 15] indicated that users are often unaware of many permissions they have already granted to their apps. Subsequently, researchers started to advocate for a more contextualized permission model that would allow users to control permissions on a more granular level [13, 34, 48].

Android adopted the runtime permission model starting in version 6.0. There are at least two reasons why runtime dialogs have the potential to improve decision making by providing context. The first is that they often (but not always) clarify to the user why a permission is needed by linking it to the functionality that is triggered, because permissions are requested at the moment the data access is needed. The second is that developers can enrich the information shown in the permission request by providing their rationale[1], which can be considered as additional contextual information. While some developers take advantage of this, many still do not.

Given that most prior results were obtained for the old permission model, it is unclear to what extent they are still applicable to the current runtime model. In this work, we conduct the first study, to the best of our knowledge, that

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

[1] `https://developer.android.com/training/permissions/requesting.html#perm-request`

examines the reasons why Android users install or remove an app *at the time this happens*, and the motivation behind granting or denying a permission *right after users make their choice*. We are also able to examine users' reasons for each permission group, thus exploring if their reasoning differs when deciding on *location*, *microphone*, *contacts*, and other types of personal data. We capture users' comfort level with their choice both at runtime as well as after the study, which allows us to compare their comfort levels with their decisions both in-context as well as after the fact. Finally, we explore whether other factors, such as demographics, may influence user decision making. Although there exist prior works that studied users' permission choices with the runtime model [29, 30, 31, 49], their goals were not focused on users' rationales.

In order to answer these questions, we employed an open-source Android app called "Paco" [11] (Personal Analytics Companion), which is a platform for ESM (Experience Sampling Method) studies. We extended Paco to be able to query users about the reasons behind the decisions they make on their Android device related to app installs, permission decisions, and uninstalling apps, and made these extensions available to the broader research community. Paco allows us to capture the rationale behind users' actions in-the-moment, when it is fresh in their minds, therefore preserving as much of the actual context as possible. The 157 participants in our study installed Paco on their personal phones and used it for a 6-week period without any interaction with us. We collected over a thousand permission decisions and the associated reasons. Our study is the first, to the best of our knowledge, to collect such data in the wild.

Our main findings include the following. Many of our participants, when deciding about permissions, are indeed thinking about whether or not the permission is needed for the app or for a specific functionality, and whether the app "should" need it. This suggests that the context provided via runtime permissions appears to be helping users make decisions. Our participants accepted 84% of all permission requests, and among those they indicated they were comfortable (right after deciding) with their choice 90% of the time. The remaining 10% of grant decisions have a low comfort score, which suggests that a form of reluctance can occur when granting permissions. When we asked participants at the end of the six week period about some of their decisions, participants were not at all comfortable with 29% of them. We also noticed that the permission denial rates vary across different permissions. For example, microphone permission requests were denied almost twice as often as storage permission requests.

We identify decision rationales for 4 events types (app installation/removal, permission grant/denial) from Android users and rank them according to participant feedback. One of the most common reasons for denying permissions was that users know they can change it later. We further break down the reasons for denials per permission type and find that the dominant rationale for each permission type can differ – sometimes significantly – across permission types.

The remainder of the paper is organized as follows. We discuss related work in Section 2, introduce our methodology in Section 3, and we detail the implementation changes to the Paco app in Section 4. Section 5 presents the results about users' rationales for app installs and removals, and Section 6 discusses the findings about permission grant and deny decisions. We summarize and discuss our findings in Section 7.

2. RELATED WORK

Existing research has explored the space of Android permissions and privacy from two perspectives, that of users and developers.

From the user perspective, research has shown that few people actually read application permission requests and even fewer comprehend them [15, 24]. In fact, users were often surprised by the abilities of background applications to collect data [23, 44], and they were concerned when presented with possible risks associated with permissions [14].

To enhance the user experience, some have suggested providing users with more privacy information and personal examples to improve comprehension [18, 25]. Researchers have designed systems to identify privacy violations and to reduce them by recommending applications based on users' security concerns [1, 10, 16, 22, 26, 50, 51, 52]. Resource requests have been categorized into benign and dangerous requests, so that only the dangerous ones require user approval, thereby reducing the number of privacy/security decisions a user needs to take [13]. Some studies employed crowdsourcing to learn user expectations and to create privacy models [28], and others explored creating personalized assistants [30].

The research focused on developer behavior has shown that many developers are not deeply knowledgeable about permissions and often misuse them [42]. Intentionally or unintentionally, they are often making mistakes [39, 40] and are not following the principle of least privilege [47]. To identify this overuse behavior, tools have been developed that employ natural language processing of application descriptions [36], and static and dynamic analysis of Android apps [3, 6, 12, 41]. Further research efforts [10, 17, 37] that design methods to generally identify malicious applications have leveraged permission overuse assessments.

To improve the situation, researchers have suggested reorganizing the permission model with better definitions and hierarchical breakdowns [5], or adding fine-grained access control for better policing [9]. A recent study by Micinski et al. suggests there should be a difference between permission accesses that happen in the background and those that happen interactively (where the access directly corresponds to a user interaction, such as when the user imports their contacts). While the former should be granted explicitly (and regularly notified to the user), the latter should be avoided to prevent user fatigue [33]. Tools have been developed that dynamically block runtime permission requests [38], or that give users the ability to deny data to applications or to substitute user data with fake data [22].

We focus on three existing pieces of research that are closest to our work. In their 2013 work on Android install-time dialogs, Kelley et al. [25] examined the extent to which the design and type of information displayed in the dialogs helps users to choose which apps to install. Both our study and theirs ask participants about factors (such as developer, popularity, reviews, etc.) that influence their choice of which app to install. Interestingly, we find different results in terms of the ranking of factors (as shown later in Section 5.2). We believe this may come from the different methods of testing, as well as the pre-Marshmallow[2] (theirs) versus post-Marshmallow (ours) permission model. A key difference between their study and ours is that they asked users to choose between pairs of apps for a friend (hypothetical scenario), whereas in our study users choose their own apps, in the wild, on their own devices.

Wijesekera et al. explored permissions in Android in two different studies [48, 49]. These studies explored how a contextualized permission model, based on the principle of *Contextual Integrity* [34] and work by Felt et al. [13], could improve dynamic permission granting. Both these studies rely on a custom version of Android 5.1.1 (pre-Marshmallow) as the study instrument, that logs every sensitive API access that requires a permission. Their first study [48] in 2015 measures how often and under what circumstances smartphone applications access protected resources regulated by permissions. They collected data on phones of 36 people about permission accesses when they happened. At the end of the week, they interviewed people, showed them screenshots of when data had been collected, and asked them if they would have liked to have prevented it (if they had been given the choice). They found that participants wanted to block 1/3 of permission accesses, citing privacy concerns over the data and lack of context about why the app needed the permission to perform its task.

In [49] the authors design a classifier to predict users permission decisions. The prediction takes into account context and generates predictions not only on-first-use, but also on subsequent uses when the context may be different. They postulate that users may not always elect to make the same decision about a permission each time it is used. They also make predictions as to when a user might change their mind, so that they do not ask on each use, but only on key ones where a user's decision may change (e.g. because of a different context). They used their predictor in a user study with 131 people and showed that they can do a far more accurate job of capturing user preferences this way than with the ask-on-first-use model. ("Ask-on-first-use" corresponds to runtime dialogs in versions of Android 6.0 or higher.) This work is very different from ours in that we do not build predictive models, and we are focused on understanding user *rationales* for decision making in the "ask-on-first-use" model. Our study also differs from all of these previous works in that we capture data "in the wild", meaning our participants used their own phones, their own choice of apps and interacted with their apps whenever they normally would.

3. METHODOLOGY

To capture users' reasoning when making privacy impacting decisions at the moment these are occurring, we use the Experience Sampling Method (ESM) [21, 27]. This method consists of asking individuals to provide systematic self-reports about their experience at random occasions during the day without the individual expecting it, often aiming to capture candid, in-the-moment experiences. Our methodology consists of surveying users at the time they are making privacy impacting decisions, by surfacing a survey when the participants install or remove an app, or when they change an app's permissions. We use the Android app Paco [11], which is part of an existing platform for ESM studies, and which can be downloaded from the Google Play store, as our study instrument.

[2] "Marshmallow" refers to Android version 6.0

In addition to the in-situ questionnaires, we ask participants to fill out an exit survey. This exit survey was used to gauge participants' privacy behaviors and technology savviness, and their awareness about permissions granted to apps on their devices. It also assesses how comfortable participants are with the permission decisions they made in the past.

Similarly to Wijesekera et al. [48], we avoid priming participants beforehand by publicizing the experiment as a study on app interactions, in order to limit response bias. No mention of privacy is made at any point during the study, except in the exit survey.

3.1 Designing the Surveys

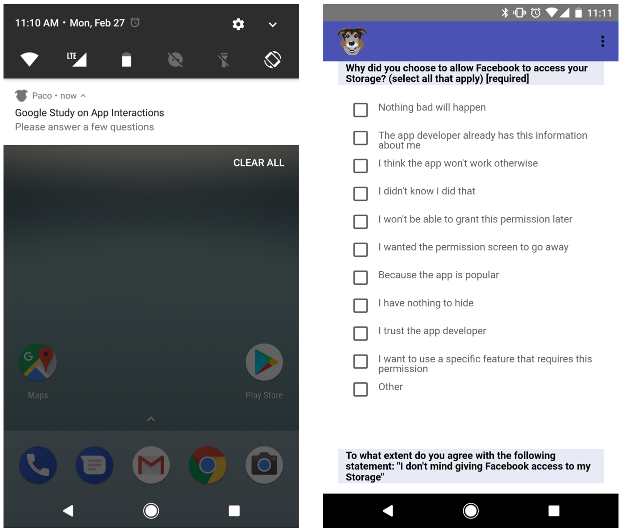
We now describe the process we followed to design our in-situ and exit surveys (provided in full in Appendix A).

3.1.1 In-Situ Surveys

The in-situ surveys are surfaced when one of the following four events occurs: the participant installs an app, removes an app, grants a permission to an app, or denies a permission to an app. In each of these cases, the participant is asked a question about his/her rationale for performing the action. In two cases, the participant receives a second question. After permission grant events, our second question aims to assess the participant's *reluctance* when allowing the permission, by asking to what extent they agree with the statement "I don't mind giving <app> access to my <permission>". App installation events also cause a second question to surface (after asking about rationales) that asks about the factors - such as app rating or popularity - that influenced their decision to install the app.

To capture the participant's decision rationales, we designed multiple-choice questions with the option to select multiple reasons, and with an additional "Other" choice allowing a free-form response. To ensure we have an exhaustive list of possible reasons, we first performed a short pre-study through the Google Surveys platform (GS), formerly known as Google Consumer Surveys (GCS). For each of the in-situ questions, we ask a random sample of 1000 participants about their reasons for performing a recent action. For instance, we asked "The last time you <did X>, what were your reasons for <doing X>?". We coded the different responses as follows. Initially two coders each coded half the responses and then cross-checked their responses. With over 90% overlap, they then independently completed the rest. The third coder independently coded responses using labels from the first two. Complete agreement was reached by all coders. Finally, we grouped answers with similar labels, and extracted the most representative answer from each of the top-10 largest groups.

Figure 1a shows how a participant is alerted that there is a question to answer, and Figure 1b shows a sample question for a permission grant request. In order to remove positional bias in the answers, we randomized the order in which the answer options were shown - with the exception of the "Other" option, which is always placed last. In order to reduce participants' response fatigue, we limit the number of questions that are surfaced to at most 3 permission events,



(a) Notification informing that a survey is available.

(b) Survey question soliciting the reasons for granting the Storage permission to an app.

Figure 1: Example of an in-situ survey in the Paco app.

2 app install events, and 1 app removal event per day, with a maximum total of 5 events per day.

3.1.2 *Exit survey*

In the exit survey, we question participants about their privacy behaviors, by asking about which privacy-enhancing practices they have employed in the past (compiled based on a Pew research survey [32]). Additionally, we ask participants to rate themselves on a 5-point scale from early to late adopters of new technology. Apart from these general questions, the exit survey also contains a personalized component. In this part, we ask participants about how comfortable they are with certain apps on their devices having access to a specific permission. These <app, permission> pairs are generated for each participant individually, by inspecting what permissions have already been granted for apps on their devices. These apps are not limited to the ones for which a permission is granted or denied during our study; they also include apps that were installed prior to enrolling in our study.

The personalized questions are worded as hypothetical scenarios, asking for example "How comfortable *would you* be with the <app name> knowing who is calling you". Moreover, the questions do not directly ask about the permissions, but rather about specific data access that this permission entails. For example, instead of asking about how comfortable the participant is with an app having storage access, we ask how comfortable they would be with the app being able to access pictures on their device. When answering such a question, participants are not informed that we selected a <app, permission> pair that exists on their devices. For each of the four permissions – Location, Contacts, Phone and Storage – we select a random app for which the permission was enabled (if available), and generate the corresponding question.

3.2 Recruitment and Incentives

Participants were recruited via our company's external U.S.-wide participant database and were sent a screening survey via email. We screened for participants using a device running Android version 6.0 or later, with their device locale set to "English - United States". (The latter requirement is needed because of the way we implemented our changes to Paco, see Section 4.2.) Participant diversity is controlled for gender, age, education and employment. Participant demographics are available in Table 1. After the recruitment phase, participants were informed that they would be required to install the Paco app. They were made aware about the fact that this app monitors their device usage to show survey questions, and were shown a list of all the data collected by Paco. Participants were told that for each of the 6 weeks they participate in our study, they would earn $10 and that submitting the exit survey would earn them an additional $20.

We recruited a total of 193 participants. Of these 193, 34 never finished the setup process and 2 voluntarily dropped out, so they are not included. The other 157 participated for the entire 6 weeks. Thirteen out of the 157 participants did not answer the exit survey, and have been excluded from parts of our analysis relying on exit survey data.

Table 1: Participant demographics

Gender	Participants	Age	Participants
Male	79	18 - 23	29
Female	78	24 - 30	44
		31 - 40	35
		41 - 50	23
		51 or over	26

Education	Participants
Up to High school	15
Some college (1-4 years, no degree)	40
Associate's degree	28
Professional school degree	5
Bachelor's degree	51
Graduate Degree	18

Employment	Participants
Arts & Entertainment	8
Business & Finance	6
Education	8
Engineering	12
Health Care	12
Human Resources	2
Information Technology	14
Management	19
Miscellaneous	15
Religion	3
Retail & Sales	17
Retired	5
Self-Employed	6
Student	18
Undisclosed	5
Unemployed	7

3.3 Ethical Considerations

In compliance with ethical training guidelines in our company, we ensured that participants' anonymity and privacy were respected. We thus carried out the following. First, all researchers have been trained in ethical user research prior to this study. Second, there was an informed consent process where the participants were informed of all the types of data being collected before they start the experiment. Third, we deleted all the participants' personally identifiable information after the data collection period and thus did not use any of it in our analysis. Fourth, respondents had the option to exit the study at any point in time. Fifth, only the data from participants who completed the entire 6 week study is used in our analysis (data from the 2 who stopped participating is discarded). Lastly, as will be explained in Section 4, we implemented end-to-end encryption on top of Paco to make sure that all gathered data would be available only to the participants and the experiment organizers (and not, for example, to operators of the Paco service or other parties).

3.4 Limitations

Our analysis is based on participant self-report data, which is subject to biases such as social desirability and recall. Participation in our study requires installing our study instrument (Paco) and enabling *accessibility* and *app usage* permissions (see Section 4.2), hence our results could be skewed towards participants willing to do so; those unwilling to do so may have characteristics we did not discover. We try to limit such an effect by recruiting a diverse participant pool (controlled for gender, age, education, and employment) and by explaining upfront about all the types of data collected. Only 2 participants, out of 193, voluntarily dropped out of the experiment expressing concerns around the accessibility permission usage, so the effect is indeed limited. In order to limit the leading effect of our in-situ questionnaire towards participants' future actions on permission decisions or app installs, we imposed upper thresholds for the number of such questionnaires, which averaged at only 30 surveyed events per user over a 6-week period.

4. TECHNICAL IMPLEMENTATION

Our main survey instrument, the Paco app [11], acts as a behavioral research platform, which allows researchers to survey participants either at predefined intervals or whenever a specific action (such as an app install or permission-related decision) occurs. The advantage of using such an app is that we do not require participants to possess a rooted Android device.

Since Paco did not provide triggers for app installation or permission change events at the time of our study, we extended its code to provide such functionality. Moreover, to ensure that the participants' data is protected while in transit between the device and our servers, we also added end-to-end encryption to Paco. All code changes to Paco were submitted and accepted to the main project, and are now available to other researchers and the general public (Paco GitHub at `https://github.com/google/paco/`).

In addition to extending the Paco platform itself, we also modify the way in which surveys are shown to the participants by making use of Paco's scripting functionality. We discuss these implementations below.

4.1 App Installation and Removal Triggers

To identify the moments when a participant installed a new app, or when they removed an app from their phone, we listen for `ACTION_PACKAGE_ADDED` and `ACTION_PACKAGE_REMOVED` intents broadcast by the Android system's package installer, while making sure that these events are not part of a package update (by checking whether the `EXTRA_REPLACING` parameter is set). For both events, we store both the package name of the app and the user-friendly app name (henceforth referred to as app name). The package name is a text string unique to each application on the Google Play store, and is useful for our analysis, whereas the app names are more identifiable and are used in generating survey questions (see Section 4.3). An example package name is `com.rovio.angrybirds` and its app name is *Angry Birds*.

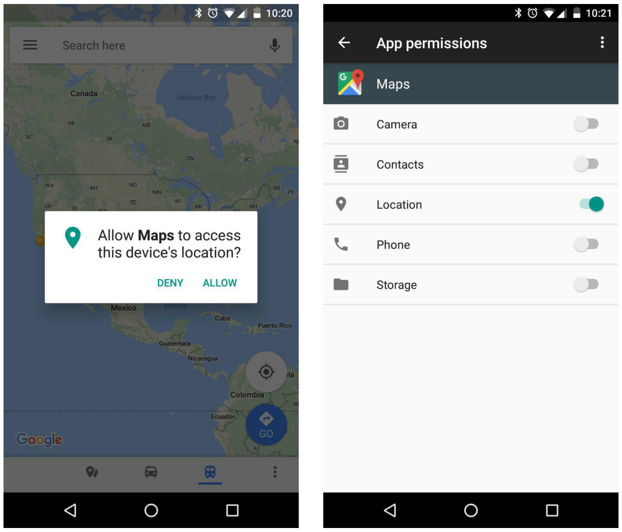
In case of an app installation event, the app name is available by querying the Android's package manager using the package name of the app. Since information about removed packages is no longer available in the package manager after an app is removed, we also manage a separate cache of package names and their corresponding app names. This allows us to access app names even after an app has been removed.

4.2 Permission Change Triggers

For permission change events, no intent is broadcast by the Android system, requiring us to monitor these permission changes ourselves. One obvious way to perform this would consist in periodically checking which permissions are granted to each of the apps installed on the user's phone, and looking for any changes in this information. This could be done by polling the Android package manager's `getInstalledPackages()` method and passing the `GET_PERMISSIONS` flag. However, a problem with this approach is that we would only detect permission *changes*, missing the case where the user has made a decision to remain in the same state as before. For instance, a user could deny a permission when it hasn't been granted before (permissions are set to deny by default when installing an app).

Because of the previous limitation, the permission change trigger is implemented as an accessibility service, which is used in Android to provide services (such as screen readers or special input devices) to people with disabilities. Because an accessibility service is able to inspect all text and user interface (UI) dialogs that are presented to the user, implementing such a service allows to analyze the text that is currently on the screen. We implement our own accessibility service to listen for events that correspond to the UI elements used for changing permissions. We then extract the text from these dialogs to determine the type of the permission and the app. We limit the accessibility service to only capture events from the `com.google.android.packageinstaller` and `com.android.settings` packages (which covers both the runtime permissions dialogs and the permission screen in the Android settings menu). This makes sure that our service does not needlessly slow down the system, and that it respects the participant's privacy by not collecting data beyond what is needed.

To identify the app for which a permission change event occurred, we query Android's usage statistics manager (this requires the *app usage* permission), determining the last active app that could have triggered a permission dialog to be shown. Because background services in Android are not allowed to request a permission, a permission dialog must always belong to the last active foreground app (if the package installer itself is excluded).



(a) The "Maps" app requesting the Location permission at runtime.

(b) Permission toggles for the "Maps" app in Android's settings.

Figure 2: Android's different methods for modifying an app's permissions.

Two different cases of permission change events are considered. The most common case is the one where an app requests a permission at runtime, either when it is first started or when the user wants to use a specific feature requiring the permission. An example of this case is depicted in Figure 2a, where the "Maps" app requests the Location permission. The second case is where the user actively changes an app's permission, by navigating through the Android's settings menus to either the screen containing all permissions for an app (see Figure 2b), or to the screen containing all apps that request a specific permission.

4.3 Generating and Surfacing Surveys

Paco allows to override the way in which surveys are generated and shown to participants, by providing experiment organizers with the ability to write scripts that will be used for generating both the notifications and the actual survey. For this study, we extensively make use of this functionality to dynamically generate questions. First, Paco's scripting functionality is used to comply with the study requirements for the in-situ questions outlined in Section 3.1.1. This includes overriding how often (and for which events) the user is notified, and randomizing the order of all survey responses except the "Other" option.

Furthermore, instead of relying on a predefined set of static questions, we generate them dynamically in order to provide more context to the participant (since the generated survey questions could be answered after a short time gap). For example, instead of asking "Why did you choose to allow the permission just now?", the participant is asked "Why did you choose to allow Maps access to your Location?".

Finally, the exit survey is also offered through Paco. This survey, too, depends heavily on dynamically generated questions. As discussed in Section 3.1, users are asked about how comfortable they are with their apps having access to data associated with a specific permission. These questions are generated for different <app, permission> pairs, where the permissions have already been granted for the app by the participant. For this purpose, the Paco app is extended with the functionality to pass on a list of all apps and their associated permissions to the script that is generating the surveys. This script selects one app for each of the four chosen permissions and generates the questions accordingly.

Table 2: Type and frequency of the different events considered by our study, and the number of events for which a participant was surveyed. See Section 3.1.1 for an explanation on survey limits.

Event Type	Occurrences	Surveyed
App Installs	3118	1913
App Removals	1944	775
Permission Grants	2239	1605
Permission Denials	437	272
Total	7738	4565

5. APP DECISIONS

5.1 Data Summary

We track four events in our study: app installs, app removals, permission grants, and permission denials. The total number of events that we recorded in our study are shown in Table 2. As mentioned in Section 3.1.1, we enforce limits on the number of events we survey each day. As a result, not all recorded events are surveyed. Our 157 participants triggered 3118 app install events (of which 1913 are surveyed), and 1944 app removals (of which 775 are surveyed). The apps could have come from either the Google Play store or from other sources. On average each participant installed 20 apps and removed 12 apps during the 6 week period. We note that a participant can install and remove the same app multiple times, and each of these actions would be recorded as a separate event. An app removal event could have occurred for an app that was installed prior to our study, and thus does not necessarily correspond to one of the app install events we observed.

We clarify that the Paco tool recorded all events (not only those surveyed) for all of the 4 event types that occurred on participants' phones during the 6 week period. Based on the complete set of user permission decisions, we observed an overall grant rate of 84% and a denial rate of 16%. Due to our self imposed limits on the number of surveys shown per day, we ended up asking survey questions for 72% of the grant events and 62% of the denial events. For the surveyed responses, we find the grant rate to be 86% (with corresponding denial rate of 14%). Thus the grant and denial rates of our surveyed (i.e., sampled) events is very close to the rates for the total occurrences. Out of the 157 participants, 144 answered the exit surveys. In the rest of the paper, we present results for the surveyed events to ensure consistency with results about participant responses.

In Figure 3, we show the activity level of our participants with our surveys. Most answered at least 10 surveys, and some have answered many more.

5.2 App Installs

After installing an app, our participants were asked to select which factors (all that apply) influenced their decision to install the app. These results are shown in Figure 4.

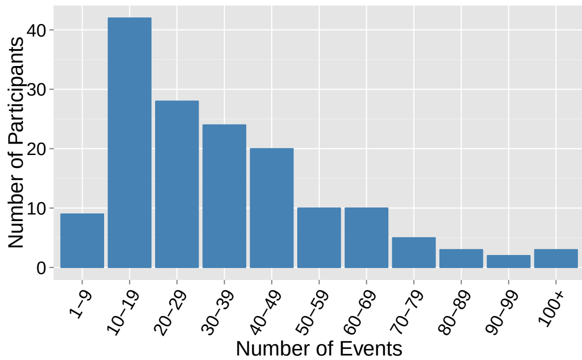


Figure 3: Event distribution across Participants

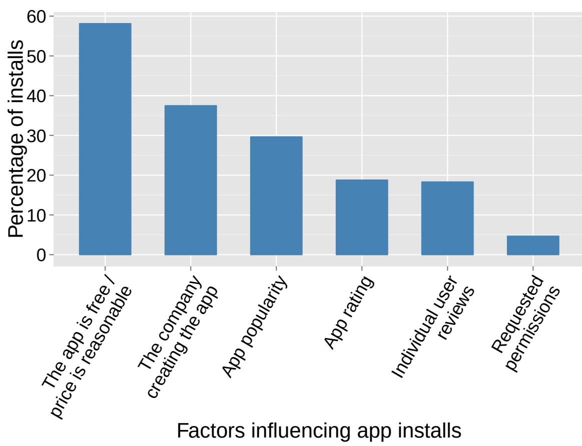


Figure 4: Factors impacting app installation (multiple responses per installation event are possible)

As expected, we observe that price is the dominant factor. What is somewhat surprising is that the company creating the app (i.e. the developer) is the second highest factor, even more important than an app's popularity. Among these six factors, permissions occur the least frequent, and only directly affect 5% of app installation decisions. This is not surprising, because with the runtime permissions model participants do not see the permission requests during the installation flow[3], and thus users are unlikely to think about permissions at that moment. However, these install events – when participants selected permissions as a factor – came from 33% of our participants; this indicates that permissions influenced one third of our participants at least once during app selection. Note that app ratings and reviews can be influenced by privacy concerns around permissions, and thus this 5% metric should actually be treated as a lower bound in terms of its ability to capture the relevance of permissions for app installation.

Our observation about the influence of permissions at installation time corroborates the finding in [25], where permissions ranked 8th out of 11 reasons. However, our findings about the influence of reviews and ratings differ significantly from those in [25] (see Figure 2 therein). They found that ratings, reviews and cost were most important (in that order) and of similar importance, whereas in our study developer and popularity were factors cited more frequently than ratings and reviews. This could be due to different study methods. They asked 366 MTurkers to rate factors on a 5-point importance scale, whereas we asked participants to select all that apply. Moreover, the MTurkers in [25] were asked about their general views, whereas our participants were asked about specific apps right after installation. This suggests that an interesting avenue for future research would be to understand if and why the influence of reviews and ratings are evolving.

[3]Some older apps that do not target an Android API level of 23 (Marshmallow) or above, and that are not yet updated to use the new permissions model, could still show a list of requested permissions at install time.

Table 3: Reasons participants checked for app installation (multiple responses per installation event are possible)

App Install Reason	Number of Occurrences (% of install events)
I want to try it out	954 (49.9%)
The app is useful	579 (30.3%)
The app is part of a product/service that I use	500 (26.1%)
The app is cool or fun to use	400 (20.9%)
I trust the app or the company making the app	310 (16.2%)
My friends/family use it	276 (14.4%)
It was the only app of its kind (no other apps provide the same functionality)	160 (8.4%)
Other	129 (6.7%)
I was required to install it	126 (6.6%)
I was offered something in return (e.g. credits, monetary rewards, discount)	79 (4.1%)
The app has fewer permissions than other apps like it	34 (1.8%)
I don't know	34 (1.8%)

Table 3 shows the reasons why users install particular apps. For each reason, the percentages indicate the proportion of install events (total events counts in Table 2) it was selected for. The reason "I want to try it out", that may capture curiosity, dominates the list and is selected in 50% of installations as a reason. The other popular reasons "The app is useful" and "The app is cool or fun to use" stress that the app's functionality plays an important role as well. We found that only 14% of the installs had social influences such as family and friends. Only 34 times (2% of the surveyed installations) did participants indicate that they compared the number of permission requests across apps before installing. However, these 34 instances originated from 15% of our participants. We hypothesize that permissions may not be a key reason at moments of installation because Android users are aware that in the runtime permissions model they can make decisions about permissions later when using the app. In Section 6.1, we see this partly confirmed since for 40% of instances when denials occurred, participants said they did so because they can grant these permissions later.

Table 4: Reasons participants checked for app removal (multiple responses per removal event are possible)

App Removal Reason	Number of Occurrences (% of removal events)
I no longer use the app	307 (39.6%)
To free up space or speed up my device	216 (27.9%)
I didn't like the app	208 (26.9%)
Other	128 (16.5%)
The app is not working as expected	120 (15.5%)
The app is crashing / very slow	48 (6.2%)
Because of advertisements in the app	42 (5.4%)
Because of in-app purchases	35 (4.5%)
The app required permissions I wasn't comfortable granting	32 (4.1%)
I don't know	16 (2.1%)

5.3 App Removals

The reasons our participants remove apps are shown in Table 4. As expected, the most common reason is that the participant no longer uses the app. The second most common reason, device performance, influenced 28% of app removals. In Section 5.1 we saw that participants are uninstalling apps at an average rate of 2/week. We were surprised by this as we assumed that when users stop using an app, they simply leave it ignored on their device rather than actively bothering to remove it. We see from these rationales that users are often removing apps for performance reasons and this contributes to the removal rate. We note that the "Other" bucket is large. Upon examination of the open ended feedback for the 128 app removal events in the "Other" option, we found that it mostly included additional details clarifying one of the already selected options. Some of the remaining responses suggested issues related to privacy or mismatched expectations. Examples include:

- Permission abuse: "The application is abusing the permission for location that I granted it. Uninstalling for this abuse of GPS." (P7)
- Negative publicity: "Read that the app is stealing private information about the phone and sending it back to China." (P31)
- Expectation mismatch: "It didn't have the information I was expecting it to have according to the description box."(P64)

Not all negative press cycles result in uninstalling apps, but for the participant above (second quote) it did. The reason "App required permissions I wasn't comfortable granting" is among the least influential here, however that option was triggered by 15% of our participants for 32 removal events. Note that if this 15% is extrapolated to the Android user base, that includes over 2 billion active devices, then the order of magnitude for devices uninstalling apps due to permissions would be in the 10s of millions.

In April 2016, the Google Play store started to require all developers to prominently disclose if their app included ads and in-app purchases. Among our participants, we see that only 10% of all uninstall events were influenced by ads or in-app purchases. This low fraction may be due to this extra transparency that helps manage people's expectations.

Table 5: Reasons participants checked for denying a permission to an app (multiple responses per deny event are possible)

Permission Deny Reason	Number of Occurrences (% of deny events)
I think the app shouldn't need this permission	111 (40.8%)
I expect the app will still work without this permission	110 (40.4%)
I can always grant it afterwards if I change my mind	110 (40.4%)
I do not use the specific feature associated with the permission	95 (34.9%)
I consider the permission to be very sensitive	57 (21%)
I don't trust the developer enough to provide this information	42 (15.4%)
I wanted the permission screen to go away	36 (13.2%)
Other	28 (10.3%)
I think something bad might happen if I provide this permission	15 (5.5%)
I didn't know I did that	7 (2.6%)
I don't know	6 (2.2%)

6. PERMISSION DECISIONS

In this section, we discuss the reasons participants provided when accepting or denying app permission requests. Our participants granted 86% of the surveyed permission requests, indicating that they were 6 times more likely to grant a permission request rather than deny it, on average. It is noteworthy that the 14% of permission requests that were denied came from 49% of our participants. This indicates that nearly half of our participants denied a permission at least once in a 6 week period. We also observed that 95% of all decisions were made via the runtime dialogs as opposed to from inside the Android settings menu. The permission grant ratio for decisions made at runtime is 86%, whereas it is only 71% for decisions made via the settings menu, implying that users are more likely to deny a permission through the settings than when deciding at runtime. One plausible explanation is that users, especially those concerned with privacy, may seek to turn off access to personal data when they are not using an app.

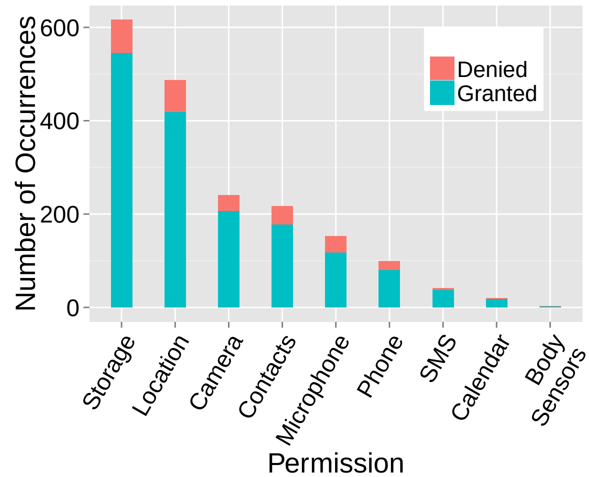
6.1 Permission denials

Table 5 shows the reasons participants had for denying permissions. Participants could pick as many reasons as they wanted for each decision, and overall the average number of reasons per denial decision was 2.3. The top two reasons imply that the majority of decisions are being made by focusing on the functionality of the app, and whether or not it really needs the particular permission. This corroborates previous findings by Wijesekera et al. [48], who observed that relevance to app functionality is a dominant reason for blocking permissions, though we find different fractions of participants who select this reason. Wijesekera et al. found that 53% of their participants wanted to block a permission because it seemed unnecessary for the app functionality. If we use our top two reasons as a proxy for their "unnecessary for app functionality" reason, our data reveals that 34% of our participants fall into this category. A potential explanation for why our study observes fewer participants denying permissions because they felt it was unnecessary is as follows. In [48] the participants were shown (at the end of the study) a handful of permission accesses that had occurred during the prior week and asked if they would have liked to deny them and why. This captures their attitude. In our study, we capture participants actions (i.e., behaviors) and their associated rationale. In essence this gap reflects a type of difference between privacy attitudes and behaviors and thus it is not surprising that the privacy behavior occurs less often than the stated attitude.

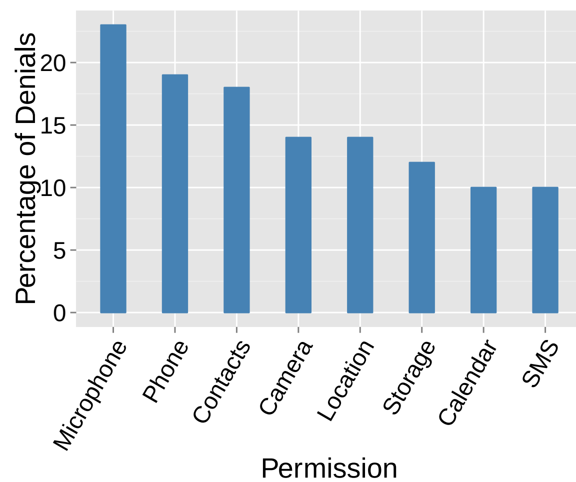
It is interesting to note that the reason "I can always grant it afterwards if I change my mind" is very prevalent among our participants (essentially tied for second place), indicating that users are aware about the fact that permissions for an app can be changed at any time (via Android's settings menu). Providing this answer for a permission denial could indicate that the user is denying the permission initially to see if the app still works, and undoing this decision later if necessary. This may indicate that the participant would prefer to use the app in a more private way and tests that possibility.

There were 57 instances where our participants denied a permission because they explicitly considered it to be very sensitive. It is striking to see that this was a more significant reason than not trusting the developer. Among these 57 instances, only 22 also picked "don't trust the developer" option. This implies that the remaining 35 instances (coming from 18 participants) correspond to scenarios were the participants do not distrust the developer but nevertheless consider the permissions sensitive and do not want to share the data. This suggests that although trust is necessary, it may not be sufficient to convince users to share data. This is of course a complex issue that requires further study because it is hard to know exactly how participants interpreted the "trust" option in our surveys.

We now examine decision making with respect to permission types. In Figure 5a, we see that the largest number of permission decisions occur for Storage and Location permissions. For each permission type, Figure 5b shows the fraction of requests that were denied. As is clear from this plot, the Microphone permission has the highest percentage of denials, followed by the Phone and Contacts permissions. It is interesting that Camera access did not exhibit a similar denial rate as Microphone; we posit that this might occur because the Camera permission sometimes only entails taking still photos (without audio and video). Although Location is perhaps the permission that users are most aware of, it does not appear among the top three most denied permissions. One possible reason is that users might have experienced some sort of habituation effect [7] for the Location permission, where a repeated exposure to such a permission



(a) Number of permission changes per permission.



(b) Percentage of permission requests denied per permission.

Figure 5: Participant Permission Decisions

request could have reduced their level of sensitivity or concern when granting such a permission, similarly to what has been reported in another study on pop-up dialogs [8].

To determine whether some decision rationales are more influential for specific permission types, we broke down our participants' reasons for permission denials according to the permission type. Figure 6 illustrates this via a heatmap. We have removed 2 permission types, SMS and Calendar, because there were fewer than 15 denials for these permissions.

Overall, we observe that the top two or three reasons for each permission type can differ. For example, for Location and Camera the top reason for denying is "I don't trust the developer". This reason has little significance for Phone and Contacts, where the dominant reasons are "I can always grant it afterwards" and "The app will still work without this permission". This shows that users make decisions about each of the permission types according to different rationales. We hypothesize that for Phone and Contacts, our participants

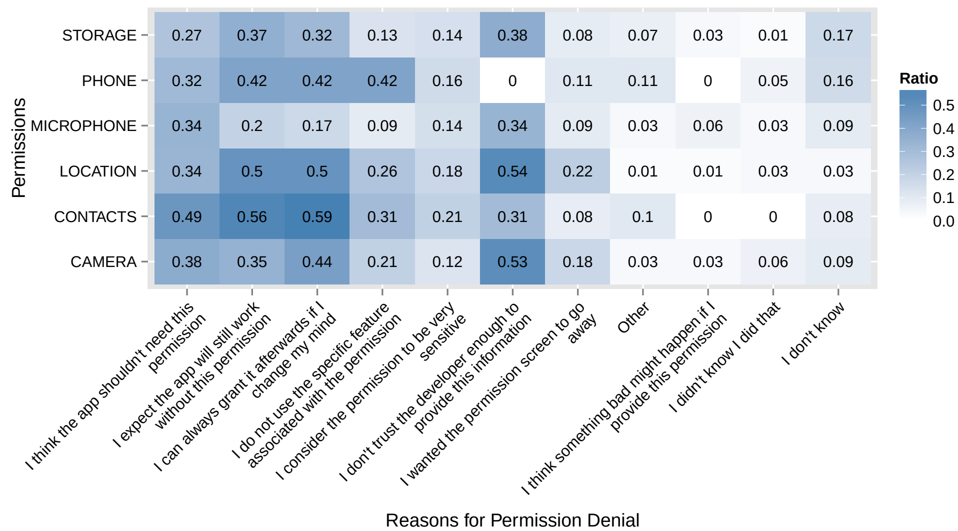


Figure 6: Reasons participants checked for denying each specific permission (multiple responses per deny event are possible). Each entry in the heatmap expresses the ratio of number of times that reason was given for the permission, over the count of all denies for the permission.

might be trying to not share them initially at all (and only doing so later if really needed) - thus issues of functionality are top of mind. However for Location and Camera, it is possible that the reason why the data is needed is often more clear and thus the primary rationale is based on trust.

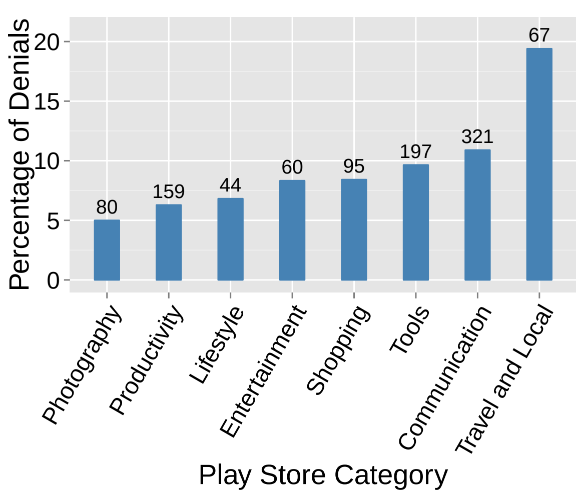


Figure 7: Percentage of permission denials across apps belonging to different Play store categories. The numbers on the bars indicate the total number of permission decisions in each category.

Next, we assess whether the permission denial rates are different across different app categories. For each of the 624 apps that registered a permission grant or denial event in our study, we identified its Play store category and considered it as an indicator of the app's functionality type. We recognize that some Play store categories, such as 'Productivity', are very broad and cover a wide range of app functionalities. However, app category was the only readily available functionality indicator.

Among the 624 apps, 41 did not appear in the Play store and seem to be device manufacturer apps that come pre-installed on the Android device or apps that have been downloaded from other Android app stores. For the remaining 583 Play apps, we aggregated the grants and denials across apps in each Play category. There were just 8 categories that had more than 20 apps, and the denial rates for these categories are shown in Figure 7. We also overlay the number of permission decisions within each category as the number on top of each bar. Denial rates vary between 5% - 19% across these 8 app categories. Moreover, the same permission can have different denial rates across different app categories. For example, 'Travel and Local' had a 43% denial rate for the Location permission, whereas 'Communication' registered only a 11% denial rate for the same permission. This reaffirms the influence of app functionality on users' permission grant or deny decisions.

6.2 Permission Grants

We now examine the reasons why users agree to grant permission requests. Table 6 shows that the dominant reason is "I want to use a specific feature that requires this permission", which suggests that users are agreeing because the request is in line with their expectations. As suggested by Felt et al. [13], a goal of using runtime dialogs is to improve the permission decision making and to avoid undermining users' expectations; our results thus indicate progress on that front. The second most important reason is trust in the developer. As discussed earlier, follow up work is necessary to fully understand how trust influences permission choices. Nonetheless, this result underscores how important it is for developers to gain a trustworthy reputation among (potential) users.

Table 6: Reasons participants checked for granting a permission to an app (multiple responses per grant event are possible)

Permission Grant Reason	Number of Occurrences (% of grant events)
I want to use a specific feature that requires this permission	1095 (68.2%)
I trust the app developer	515 (32.1%)
I think the app won't work otherwise	382 (23.8%)
I have nothing to hide	289 (18%)
Nothing bad will happen	225 (14%)
The app developer already has this information about me	208 (13%)
I wanted the permission screen to go away	164 (10.2%)
Because the app is popular	150 (9.3%)
Other	39 (2.4%)
I didn't know I did that	36 (2.2%)
I won't be able to grant this permission later	22 (1.4%)

In a similar way as we did for the denials case, we checked whether some reasons are more influential for specific permission types, but found the distribution of reasons to be similar across permission types.

Next we look at the question of whether or not participants grant permissions willingly. Recall that after our participants granted a permission, we asked them to indicate if they agree or disagree (5 pt scale) with the statement "I don't mind giving <app> access to my <permission>" (Q2 in Appendix A.1.3). Surprisingly, we found that 10% of the time, participants indicated that they "Disagree" or "Strongly disagree" with the statement (see Figure 8). This could occur if participants believe an app won't work without the requested permission and so they agree, albeit reluctantly. This can be associated with the phenomenon of "learned helplessness" [46], which covers scenarios when participants convince themselves they agree with something (e.g., data sharing) because they did not really have a choice.

To see whether this comfort level changes over time, we asked participants in the exit survey to rate their comfort level with permissions they had granted to apps on their phones in the past (Q19 – Q22 in Section A.2; we included "I don't know the app" as an additional option). When asking these questions, we made the permissions more specific. For example, if the participant had granted the Storage permission, we ask whether they were comfortable with the app accessing photos on their device storage. These questions were intended not only to revisit comfort with prior decisions, but also to illustrate more explicitly to the participants the implication of their decision. These prior decisions may have occurred any time during our 6 week study or even earlier as explained in Section 3.1.2.

In a surprisingly high number of situations (see Figure 9) participants were not comfortable with their prior decisions. In 29% of scenarios presented to the participants, they indicated they were "Not at all comfortable" with the data access that was allowed to the app. If we include the cases where users were "Slightly comfortable", then we see that in 44% of the cases our participants are not feeling comfortable about their past decisions. These discomfort levels vary based on the permission: on a scale from 1 to 5, where larger numbers indicate a higher discomfort, the Storage permission entails an average discomfort of 3.41, Phone has a discomfort of 3.33, Contacts has a discomfort 3.11, and Location has a discomfort of 2.77.

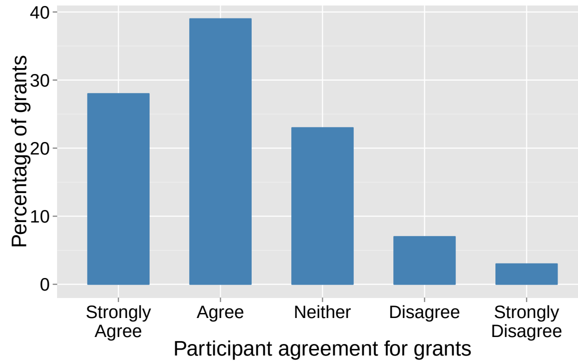


Figure 8: Participant responses to the statement: "I don't mind giving <app> access to my <permission>", right after granting that permission.

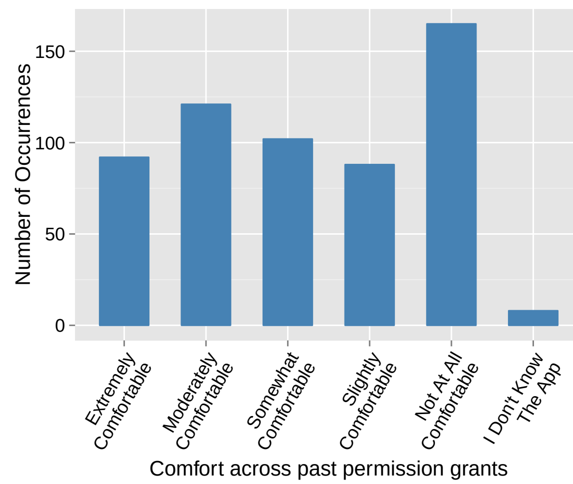


Figure 9: Participant comfort for permissions that were granted in the past, in response to the exit survey question "How comfortable would you be with the <app name> app knowing <information available through the permission>".

Participants were not comfortable about permissions they granted in the past and this may be occurring because they do not always understand what a permission entails, and only realize this after it is made explicit. Consider, for example, the Storage permission: this permission might be understood by a user as allowing the app to store data on the device, only to be refuted by our question stating that the app now has access to pictures on the user's device. This explanation is supported by previous work [18, 43] that has shown how users need to be confronted with a specific scenario before being able to correctly reason about privacy and security.

It is interesting to contrast the 29% discomfort long after decision making, to the 10% reluctance that existed at the moment of decision making. This 29% statistic could be said to capture privacy attitudes; the exit survey captures what people say or think about sharing data when they are being questioned but not making a real life decision. However in practice, in only 10% of grant decisions did users say that they minded sharing the data right after granting. The gap between these numbers approximately captures the difference in participant's attitudes and behaviors, in the context of Android permissions.

6.3 Other influences

We check whether the participants' demographics are associated with their grant/denial behavior. We used Pearson's Chi-squared test (with Yates' continuity correction when needed) to check the dependence between participants' age and gender, and their denial behavior. We control for age (gender) when gender (age) is being tested. Due to small sample sizes, we did not test for independence across education and employment demographics. We notice that women across age groups 18-23 ($\tilde{\chi}^2 = 10.7$, df = 1, p-value = 0.001068) and 31-40 ($\tilde{\chi}^2 = 16.3$, df = 1, p-value = 5.396e-05) are three times as likely to deny permissions than men. On average over all age groups, women deny twice as often as men, with a 20% denial rate for women compared to 11% for men ($\tilde{\chi}^2 = 25.6$, df = 1, p-value = 4.11e-07). Comparing men across different age groups, we notice that men's denial rates differ significantly ($\tilde{\chi}^2 = 31.2$, df = 4, p-value = 2.841e-06); participants in age ranges 18-23 and 31-40 have denial rates around 5% whereas the other age groups have denial rates of 15% or higher, about three times higher.

Lastly, we checked associations between participant responses to questions in the exit survey (Q1–Q18 in Section A.2) and their permission denials. We did not find any statistically significant correlations or dependencies.

7. DISCUSSION AND CONCLUSION

There are a couple of important takeaways herein for Play store developers. First, we saw that in terms of app installs and uninstalls, permissions were not a dominant reason compared to other reasons. However, 15% of our participants uninstalled apps due to permissions. Extrapolating this statistic to the set of Android devices (over 2 billion), indicates that this could affect tens of millions of devices. This result could motivate developers to reconsider requesting certain permissions at all or to make runtime requests more contextual – for example by only asking for permission access when the user opts to use certain functionality within their app rather than at first run.

Second, the vast majority of rationales for decision making around permissions are related to app functionality, whether the app needs the permission, whether it "should" need it, and whether the user needs the functionality entailed by it. Thus, participants are more willing to grant permissions when the reason for asking is clear. This should motivate developers to provide sufficient and clear explanations for their requests. Android provides a utility method (`shouldShowRequestPermissionRationale()`) to help identify situations where users might need an explanation.

In summary, we observed an overall denial rate of 16%. These denies came from half our participants which indicates that there exists one or some scenarios for many people in which they will deny a permission. The scenarios when participants deny permissions are very varied. This is implied by the findings that i) denial rates vary from 10% to 23% according to permission type, and ii) denial rates vary from 5% to 19% across app genres (Play store categories). Among our participants, we also saw that women denied permissions roughly twice as often as men.

We found that even though the overall grant rate is quite high, at 84%, there is a portion of decisions (10%) in which users grant permissions reluctantly. Moreover, users were surprisingly uncomfortable (29%) when revisiting their prior decisions at the end of our study. This indicates a gap between behaviors and stated attitudes.

Our participants' rationale for denying a permission in 42% of denial instances, was because they knew they could change the permissions afterwards. We hypothesize that this might be happening because participants want to test out whether or not the app will work in a more privacy preserving way (with less user data). Exploring this would be an interesting avenue for future research.

It is interesting albeit hard to understand how users' comfort levels and understanding of permissions have evolved after the introduction of runtime dialogs. In [48] (pre-runtime), the authors state that 80% of their participants wanted to deny at least one permission. In our study, we recorded that 49% of our participants denied permissions at least once. We found that 16% of permission requests were denied. This is about half the rate reported in [48], though the latter study asked participants to allow or deny access many times a permission was used, instead of only on first use as in our study. These two studies differ in their interactions with users, and both involve limited populations, yet these metrics hint that users may be getting more comfortable granting permissions using runtime dialogs. It would be interesting to explore this hypothesis in future research that makes a more direct comparison.

8. ACKNOWLEDGMENTS

Many thanks to Bob Evans for collaborating with us on extending Paco; to Rob Reeder, Sunny Consolvo, Tara Matthews, Allison Woodruff, Jeffrey Warshaw, and Manya Sleeper for valuable suggestions on the survey design; to Patrick Gage Kelley for providing feedback on an initial draft; to Clara Sherley-Appel for helping with participant onboarding; to Svetoslav Ganov and Giles Hogben for suggestions on technical implementation.

9. REFERENCES

[1] H. M. Almohri, D. D. Yao, and D. Kafura. Droidbarrier: Know what is executing on your android. In *Proceedings of the 4th ACM Conference on Data and Application Security and Privacy*, CODASPY. ACM, 2014.

[2] H. Almuhimedi, F. Schaub, N. Sadeh, I. Adjerid, A. Acquisti, J. Gluck, L. F. Cranor, and Y. 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. ACM, 2015.

[3] K. W. Y. Au, Y. F. Zhou, Z. Huang, and D. Lie. Pscout: Analyzing the android permission specification. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security*, CCS. ACM, 2012.

[4] R. Balebako, J. Jung, W. Lu, L. F. Cranor, and C. Nguyen. Little brothers watching you: Raising awareness of data leaks on smartphones. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS. ACM, 2013.

[5] D. Barrera, H. G. Kayacik, P. C. van Oorschot, and A. Somayaji. A methodology for empirical analysis of permission-based security models and its application to android. In *Proceedings of the 17th ACM Conference on Computer and Communications Security*, CCS. ACM, 2010.

[6] E. Bodden. Easily instrumenting android applications for security purposes. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, CCS. ACM, 2013.

[7] M. E. Bouton. *Learning and behavior: A contemporary synthesis.* Sinauer Associates, 2007.

[8] C. Bravo-Lillo, L. Cranor, S. Komanduri, S. Schechter, and M. Sleeper. Harder to ignore? revisiting pop-up fatigue and approaches to prevent it. In *10th Symposium On Usable Privacy and Security*, SOUPS. USENIX Association, 2014.

[9] S. Bugiel, S. Heuser, and A.-R. Sadeghi. Flexible and fine-grained mandatory access control on android for diverse security and privacy policies. In *Proceedings of the 22Nd USENIX Conference on Security*, SEC. USENIX Association, 2013.

[10] W. Enck, P. Gilbert, B.-G. 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 *Proceedings of the 9th USENIX Conference on Operating Systems Design and Implementation*, OSDI. USENIX Association, 2010.

[11] B. Evans. Paco – applying computational methods to scale qualitative methods. In *Ethnographic Praxis in Industry Conference Proceedings*. Wiley Online Library, 2016.

[12] A. P. Felt, E. Chin, S. Hanna, D. Song, and D. Wagner. Android permissions demystified. In *Proceedings of the 18th ACM Conference on Computer and Communications Security*, CCS. ACM, 2011.

[13] A. P. Felt, S. Egelman, M. Finifter, D. Akhawe, and D. Wagner. How to ask for permission. In *Proceedings of 7th Usenix conference on Hot Topics in Security (HotSec)*, 2012.

[14] A. P. Felt, S. Egelman, and D. Wagner. I've got 99 problems, but vibration ain't one: A survey of smartphone users' concerns. In *Proceedings of the Second ACM Workshop on Security and Privacy in Smartphones and Mobile Devices*, SPSM. ACM, 2012.

[15] A. P. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D. Wagner. Android permissions: User attention, comprehension, and behavior. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS. ACM, 2012.

[16] C. Gibler, J. Crussell, J. Erickson, and H. Chen. Androidleaks: automatically detecting potential privacy leaks in android applications on a large scale. In *International Conference on Trust and Trustworthy Computing*. Springer, 2012.

[17] A. Gorla, I. Tavecchia, F. Gross, and A. Zeller. Checking app behavior against app descriptions. In *Proceedings of the 36th International Conference on Software Engineering*, ICSE. ACM, 2014.

[18] M. Harbach, M. Hettig, S. Weber, and M. Smith. Using personal examples to improve risk communication for security & privacy decisions. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI. ACM, 2014.

[19] M. A. Harris, R. Brookshire, and A. G. Chin. Identifying factors influencing consumers' intent to install mobile applications. *International Journal of Information Management*, 2016.

[20] M. A. Harris, R. Brookshire, K. Patten, and B. Regan. Mobile application installation influences: have mobile device users become desensitized to excessive permission requests? In *Proceedings of the Twentieth Americas Conference on Information Systems*, AMCIS, 2015.

[21] S. E. Hormuth. The sampling of experiences in situ. *Journal of Personality*, 1986.

[22] P. Hornyack, S. Han, J. Jung, S. Schechter, and D. Wetherall. These aren't the droids you're looking for: Retrofitting android to protect data from imperious applications. In *Proceedings of the 18th ACM Conference on Computer and Communications Security*, CCS. ACM, 2011.

[23] J. Jung, S. Han, and D. Wetherall. Short paper: Enhancing mobile application permissions with runtime feedback and constraints. In *Proceedings of the Second ACM Workshop on Security and Privacy in Smartphones and Mobile Devices*, SPSM. ACM, 2012.

[24] P. G. Kelley, S. Consolvo, L. F. Cranor, J. Jung, N. Sadeh, and D. Wetherall. A conundrum of permissions: Installing applications on an android smartphone. In *Proceedings of the 16th International Conference on Financial Cryptography and Data Security*, FC. Springer-Verlag, 2012.

[25] P. G. Kelley, L. F. Cranor, and N. Sadeh. Privacy as part of the app decision-making process. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI. ACM, 2013.

[26] W. Klieber, L. Flynn, A. Bhosale, L. Jia, and L. Bauer. Android taint flow analysis for app sets. In *Proceedings of the 3rd ACM SIGPLAN International Workshop on the State of the Art in Java Program Analysis*, SOAP. ACM, 2014.

[27] R. Larson and M. Csikszentmihalyi. The experience sampling method. *New Directions for Methodology of Social & Behavioral Science*, 1983.

[28] 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, UbiComp. ACM, 2012.

[29] J. Lin, B. Liu, N. Sadeh, and J. I. Hong. Modeling users' mobile app privacy preferences: Restoring usability in a sea of permission settings. In *Symposium on Usable Privacy and Security*, SOUPS, 2014.

[30] B. Liu, M. S. Andersen, F. Schaub, H. Almuhimedi, S. A. Zhang, N. Sadeh, Y. Agarwal, and A. Acquisti. Follow my recommendations: A personalized privacy assistant for mobile app permissions. In *Twelfth Symposium on Usable Privacy and Security*, SOUPS. USENIX Association, 2016.

[31] B. Liu, J. Lin, and N. Sadeh. Reconciling mobile app privacy and usability on smartphones: Could user privacy profiles help? In *Proceedings of the 23rd International Conference on World wide web*, WWW. ACM, 2014.

[32] M. Madden and L. Rainie. Americans' Attitudes About Privacy, Security and Surveillance. http://www.pewinternet.org/2015/05/20/americans-attitudes-about-privacy-security-and-surveillance/.

[33] K. Micinski, D. Votipka, R. Stevens, N. Kofinas, M. L. Mazurek, and J. S. Foster. User interactions and permission use on android. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI. ACM, 2017.

[34] H. Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 2004.

[35] P. A. Norberg, D. R. Horne, and D. A. Horne. The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs*, 2007.

[36] R. Pandita, X. Xiao, W. Yang, W. Enck, and T. Xie. Whyper: Towards automating risk assessment of mobile applications. In *Proceedings of the 22Nd USENIX Conference on Security*, SEC. USENIX Association, 2013.

[37] B. P. Sarma, N. Li, C. Gates, R. Potharaju, C. Nita-Rotaru, and I. Molloy. Android permissions: A perspective combining risks and benefits. In *Proceedings of the 17th ACM Symposium on Access Control Models and Technologies*, SACMAT. ACM, 2012.

[38] B. Shebaro, O. Oluwatimi, D. Midi, and E. Bertino. Identidroid: Android can finally wear its anonymous suit. *Transactions on Data Privacy*, 2014.

[39] I. Shklovski, S. D. Mainwaring, H. H. Skúladóttir, and H. 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. ACM, 2014.

[40] M. Smith. Usable Security – The Source Awakens. *Usenix Enigma*, 2016.

[41] M. Spreitzenbarth, F. Freiling, F. Echtler, T. Schreck, and J. Hoffmann. Mobile-sandbox: Having a deeper look into android applications. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC. ACM, 2013.

[42] R. Stevens, J. Ganz, V. Filkov, P. Devanbu, and H. Chen. Asking for (and about) permissions used by android apps. In *2013 10th Working Conference on Mining Software Repositories (MSR)*, 2013.

[43] C. Swanson, R. Urner, and E. Lank. Naïve security in a wi-fi world. In *IFIP International Conference on Trust Management*. Springer, 2010.

[44] C. Thompson, M. Johnson, S. Egelman, D. Wagner, and J. King. When it's better to ask forgiveness than get permission: Attribution mechanisms for smartphone resources. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS. ACM, 2013.

[45] A. Vaidos. Google Play Store vs Apple's App Store - A Comparison. http://news.softpedia.com/news/google-play-store-vs-apple-s-app-store-a-comparison-512601.shtml.

[46] J. Warshaw, T. Matthews, S. Whittaker, C. Kau, M. Bengualid, and B. A. Smith. Can an algorithm know the "real you"?: Understanding people's reactions to hyper-personal analytics systems. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI. ACM, 2015.

[47] X. Wei, L. Gomez, I. Neamtiu, and M. Faloutsos. Permission evolution in the android ecosystem. In *Proceedings of the 28th Annual Computer Security Applications Conference*, ACSAC. ACM, 2012.

[48] P. Wijesekera, A. Baokar, A. Hosseini, S. Egelman, D. Wagner, and K. Beznosov. Android permissions remystified: A field study on contextual integrity. In *Proceedings of the 24th USENIX Conference on Security Symposium*, SEC. USENIX Association, 2015.

[49] P. Wijesekera, A. Baokar, L. Tsai, J. Reardon, S. Egelman, D. Wagner, and K. Beznosov. The feasibility of dynamically granted permissions: Aligning mobile privacy with user preferences. In *Proceedings of the 38th IEEE Symposium on Security and Privacy*. IEEE, 2017.

[50] R. Xu, H. Saïdi, and R. Anderson. Aurasium: Practical policy enforcement for android applications. In *Proceedings of the 21st USENIX Conference on Security Symposium*, Security. USENIX Association, 2012.

[51] Y. Zhang, M. Yang, B. Xu, Z. Yang, G. Gu, P. Ning, X. S. Wang, and B. Zang. Vetting undesirable behaviors in android apps with permission use analysis. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, CCS. ACM, 2013.

[52] H. Zhu, H. Xiong, Y. Ge, and E. Chen. Mobile app recommendations with security and privacy awareness. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD. ACM, 2014.

APPENDIX

A. SURVEY QUESTIONS

Responses to all questions are required.

A.1 In-situ questions

A.1.1 App installation scenario

The order of possible responses to the questions for the in-situ survey is always randomized (with the exception of the 'Other' option, which is always placed last).

Q1: Which factors influenced your decision to install <app>? (select all that apply)

- App rating
- App popularity
- Individual user reviews
- Requested permissions
- The company creating the app
- The app is free / price is reasonable

Q2: Why did you install <app>? (select all that apply)

- The app has fewer permissions than other apps like it
- My friends/family use it
- I want to try it out
- I was required to install it
- The app is part of a product/service that I use
- The app is useful
- The app is cool or fun to use
- I trust the app or the company making the app
- It was the only app of its kind (no other apps provide the same functionality)
- I was offered something in return (e.g. credits, monetary rewards, discount)
- I don't know
- Other: ________

A.1.2 App removal scenario

Q1: Why did you remove <app>? (select all that apply)

- The app required permissions I wasn't comfortable with granting
- I no longer use the app
- To free up space or speed up my device
- Because of advertisements in the app
- Because of in-app purchases
- I didn't like the app
- The app is crashing / very slow
- The app is not working as expected
- I don't know
- Other: ________

A.1.3 Permission grant scenario

Q1: Why did you choose to allow <app> to access your <permission>? (select all that apply)

- I want to use a specific feature that requires this permission
- I think the app won't work otherwise
- I trust the app developer
- Because the app is popular
- I won't be able to grant this permission later
- I have nothing to hide
- I wanted the permission screen to go away
- Nothing bad will happen
- I didn't know I did that
- I don't know
- The app developer already has this information about me
- Other: ________

Q2: To what extent do you agree with the following statement: "I don't mind giving <app> access to my <permission>"?

- Strongly disagree
- Disagree
- Neither agree or disagree
- Agree
- Strongly agree

A.1.4 Permission deny scenario

Q1: Why did you deny <app> to have access to your <permission>? (select all that apply)

- I do not use the specific feature associated with the permission
- I think the app shouldn't need this permission
- I expect the app will still work without this permission
- I consider the permission to be very sensitive
- I don't trust the developer enough to provide this information
- I can always grant it afterwards if I change my mind
- I wanted the permission screen to go away
- I think something bad might happen if I provide this permission
- I don't know
- I didn't know I did that
- Other: ________

A.2 Exit Survey

Each of the questions Q1-Q15 have the same three possible answers:

- Yes
- No
- I don't know what this is / means

Q1: Have you ever blocked another person on a social network?

Q2: Have you ever deleted an online account?

Q3: Have you ever downloaded your historical data from an account (e.g. Google Takeout)?

Q4: Have you ever changed the privacy settings for any of your accounts?

Q5: Have you ever read part or all of an online privacy policy?

Q6: Have you ever decided not to install an app on your mobile device because of permissions it requested?

Q7: Have you ever uninstalled an app on your mobile device because of permissions it used?

Q8: Have you ever declined to give an app permission to do something on your mobile device?

Q9: Have you ever declined to use a website because it asked for information you did not want to provide?

Q10: Have you ever stopped using an Internet service or website because you were concerned about how it might use your personal information?

Q11: Have you ever cleared cookies and/or browser history?

Q12: Have you ever installed software to block ads?

Q13: Have you ever installed software to stop websites from tracking what you do online?

Q14: Have you ever used a password manager?

Q15: Have you ever used account settings to limit the data that could be collected or used?

Q16: Which of the following best describes the time at which you try new technology?

- As soon as the technology is available / among the first people to try it
- Sooner than most people, but not among the first
- Once many people are using it
- Once most people are using it
- I don't usually buy or try out new technology

Q17: When an Internet company collects data about you while you are online, overall how beneficial or harmful is that likely to be for you?

- Extremely beneficial
- Moderately beneficial
- Slightly beneficial
- Neither beneficial nor harmful
- Slightly harmful
- Moderately harmful
- Extremely harmful

In questions Q18-Q22, we used a 5-pt Likert scale to measure comfort.

Q18: How comfortable or uncomfortable are you with online companies collecting data about what you do online?

- Extremely Comfortable
- Moderately Comfortable
- Somewhat Comfortable
- Slightly Comfortable
- Not at all Comfortable

In addition to the 5-pt comfort scale, for questions Q19-Q22 users could also select an option "I don't know the app" if they do not recognize the app in the question. The apps we showed users were ones on their phones, so most of the time apps should be recognized.

Q19: How comfortable would you be with the <app name> app knowing your home and work address? (only surfaced if an app exists that was given the Location permission)

- Extremely Comfortable
- Moderately Comfortable
- Somewhat Comfortable
- Slightly Comfortable
- Not at all Comfortable
- I don't know the app

The question answer options for Q20-Q22, were the same as in Q19.

Q20: How comfortable would you be with the <app name> app knowing the phone numbers of your friends and family? (only surfaced if an app exists that was given the Contacts permission)

Q21: How comfortable would you be with the <app name> app knowing who is calling you? (only surfaced if an app exists that was given the Phone permission)

Q22: How comfortable would you be with the <app name> app seeing the pictures taken with your camera? (only surfaced if an app exists that was given the Storage permission)

Q23: Do you have any feedback for us? Is there anything else you would like to tell us?
Open ended response

Weighing Context and Trade-offs: How Suburban Adults Selected Their Online Security Posture

Scott Ruoti[‡†*], Tyler Monson[‡], Justin Wu[‡], Daniel Zappala[‡], Kent Seamons[‡]
Brigham Young University[‡]
Sandia National Laboratories[†]
scott@ruoti.org, *monson@isrl.byu.edu*, *justinwu@byu.edu*, *zappala@cs.byu.edu*, *seamons@cs.byu.edu*

ABSTRACT

Understanding how people behave when faced with complex security situations is essential to designing usable security tools. To better understand users' perceptions of their digital lives and how they managed their online security posture, we conducted a series of 23 semi-structured interviews with mostly middle-aged parents from suburban Washington state. Using a grounded theory methodology, we analyzed the interview data and found that participants chose their security posture based on the immense value the Internet provides and their belief that no combination of technology could make them perfectly safe. Within this context, users have a four-stage process for determining which security measures to adopt: learning, evaluation of risks, estimation of impact, and weighing trade-offs to various coping strategies. Our results also revealed that a majority of participants understand the basic principles of symmetric encryption. We found that participants' misconceptions related to browser-based TLS indicators lead to insecure behavior, and it is the permanence of encrypted email that causes participants to doubt that it is secure. We conclude with a discussion of possible responses to this research and avenues for future research.

1. INTRODUCTION

Security has been a persistent problem for the Internet; attacks against corporations [7, 16, 24, 29, 32, 40] and individuals [36, 19, 26] are now commonplace. The literature is rife with recommendations and tools (e.g., password managers, secure email) from security experts for improving users' security postures [6, 8, 39, 34]. Unfortunately, users are slow to adopt these practices, leading them to fall victim to the same categories of attack that have been pervasive for over a decade (e.g., weak passwords, phishing).

To address this problem, it is important to ask why users reject this advice, as the answer to this question should guide the direction of future research. If users are unaware of available protections, then the community needs to research how to best disseminate this knowledge. If users do not want to be bothered with security, then research should focus on technologies that act without user input or awareness. If instead, users reject security advice because it is too costly to implement (e.g., time, effort, money), then we need to better understand users' internal models of security and design protections that fit within that context.

We conducted a grounded theory study [9] on how users perceive their digital lives and how they manage their online security posture. As part of this effort, we conducted a series of 23 semi-structured interviews with mostly middle-aged parents in a suburban location in Washington state. While participants were free to self-guide the interview, the following topics were discussed: (a) their awareness of the potential risks associated with their online activity, (b) which risks they actively mitigated, (c) the steps they took to mitigate those risks, and (d) why they chose not to mitigate others. To explore how they viewed specific security contexts, we asked participants about their understanding and opinions regarding various security technologies associated with the web (e.g., encryption, TLS, secure messaging).

Our analysis revealed that the context within which participants select their security posture is dominated by two key factors. First, the perception that the Internet has brought incredible value into their lives, and most limitations on its usage would be extremely damaging. Second, the perception that regardless of what steps are taken, they can never be perfectly safe, which curbs any desire to implement security mechanisms that carry a high cost of adoption. Because perfect security is perceived as unattainable, users instead engage in a four-step process wherein they weigh the costs and benefits of various coping strategies designed to minimize the likelihood or impact of online risks against the benefits they derive from online activity.

1. **A user learns about a new security threat.** This happens by word of mouth, news reports, television shows, and movies.

2. **The user evaluates the risk presented by the threat.** If the attack seems sufficiently unlikely, they will generally ignore it.

3. **The user estimates the impact of a successful attack.** The amount of damage is commensurate to the effort they are willing to expend to address the threat.

[*] Scott Ruoti now works at MIT Lincoln Laboratory

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

4. **The user selects an appropriate coping strategy.** This selection is based on trade-offs between the cost of implementing the coping strategy (e.g., diminished ability to use the Internet) and its ability to mitigate risk by reducing attack surface and/or impact.

Importantly, users are fluid in their application of this process and do not necessarily proceed linearly through a series of steps. Rather, they may skip some steps or re-evaluate past steps as they learn new information.

As part of our grounded theory methodology, we avoided investigating related work before completing our analysis of participants' responses. This was done to avoid biasing ourselves as we designed, administered, and analyzed our study, allowing us to focus on what the data was saying, and not what prior research had found. After reviewing the related work, we found that the above process has a strong relationship to the inputs and outputs of protection motivation theory [33]. Our work is useful in demonstrating how users adapt this model to online activity and also extends upon this model by describing how users weigh trade-offs when selecting coping strategies.

Our analysis of the data also revealed several other topics that were particularly interesting:

- *Most participants understand the basic principles of symmetric key cryptography.* They correctly identified that encryption relies on a shared key, and only owners of this key could read an encrypted message.
- *Participants' belief that TLS indicators represented site safety, not connection security, led them to click through TLS connection warning pages.* More troubling, they were most likely to ignore the warning pages for well-known sites (e.g., Amazon, Microsoft, Google) when, in reality, warnings on these sites are relatively more likely to indicate malicious behavior.
- *Participants felt that secure email was less secure than texting because of its permanence.* In line with their views that nothing is 100% safe, permanence meant that at any time in the future an attacker (e.g., government, hacker) could choose to break their old email, whereas text messages were viewed as ephemeral and only vulnerable to an active wire-tapper.

2. RELATED WORK

There is a large body of literature that relates to understanding user motivation, perception, and behavior in the context of security. We first discuss general theories of user behavior and then examine relevant work in the usable security field that relates to the perception of risk, cost-benefit tradeoffs, user motivation, and experience with security warnings.

2.1 Theories of User Behavior

Numerous theories have been developed by psychologists regarding how users can be persuaded to take some action, such as adopting health advice or purchasing a product [18]. Several of these have been used to study persuasion in the context of security and privacy behaviors. For example, the elaboration likelihood model states that there is a central route to persuasion, in which a person carefully considers the merits of information presented, and a peripheral route that involves positive and negative cues [30]. For example, this model has been applied to understand adoption of electronic health records [4] and trust in online retailers [45].

Protection motivation theory states that people react to fears by assessing the severity and probability of the threat and then appraising the efficacy of a recommended behavior and their ability to carry out that recommendation effectively [33]. This theory has been used to explain home computer user's security behavior [3], the use of anti-virus software [28], and the effectiveness of security policies in the workplace [21]. LaRose et al. [27] use both of these theories, along with social cognitive theory, to develop a framework for motivating safe behavior online.

Witte developed the extended parallel process model (EPPM) to explain how people react when confronted with communications that appeal to fear [44]. In EPPM, user reactions to threats are driven by the assessment of a threat and efficacy, and their reaction is either determined by fear control or danger control. If there is a perception of high threat and high efficacy, then people will take the appropriate protective action (danger control). However, if there is high threat and low efficacy, people will lose hope and reject the proposed remedy (fear control). Based on this theory, appeals to take protective security measures need to ensure that people respond more strongly to the effectiveness of the proposed remedy and their capability to implement it than to the fear of the threat. Too strong of an appeal to fear leads to inaction.

Our theory is most similar to EPPM, with overlap in the concepts that people appraise the risk and severity of a threat, as well as the use of cognitive defense mechanisms to manage anxiety. Many users have internalized a fear that nothing is safe on the Internet, but this fear has generally not been strong enough to override the belief that the Internet nonetheless offers significant value. However, users may choose to avoid certain activities if their fear is too strong. Extending this model, we find that users weigh cost-benefit trade-offs in their evaluation of response- and self-efficacy. This is similar to work by Herley [22], which argues that users' rejection of some security advice is rational from an economic perspective. He discusses how in the context of password composition, phishing, and TLS warnings users have an economic incentive to ignore security advice; the cost of addressing these issues is greater than the reduction in harm. While Herley's work was theoretical, our study grounds his ideas in data, demonstrating that users do think through these economic arguments—though, in a simpler form—when deciding what security decisions to make.

2.2 Risk Perception and Behavior

Other work in usable security has examined users' perception of risk and how this motivates behavior. Wash interviewed participants regarding their perceptions of digital security [41] and identified eight "folk models" describing participants' understanding of viruses, malware, and hackers. Wash also discussed how these models could explain why participants ignored security warnings. This paper has many similarities to our work, using a similar methodology and population. While both works discuss online threats, our work provides more details regarding the harm that users associate with risk and the context under which users make security decisions.

Wash and Rader have also studied security beliefs and how this affects how people choose to protect their home computer [42]. They find that direct and visible threats lead to positive security decisions, while beliefs that require more technical knowledge lead to fewer precautions. The educated and older are more likely to hold these more sophisticated beliefs. We confirm these finding in our research, finding that users feel overly-technical solutions often offer marginal benefits in comparison to their adoption cost.

Harbach et al. surveyed users and asked them what risks they were most concerned about for five different online scenarios [20]. In addition to stating potential risks, participants were also asked to rank them. Finally, participants were presented with a list of 22 common risks and asked to rate how relevant they found those risks. They found that users were aware of far fewer risks than had previously been believed and recommended that more work be done in risk communication and education. Contrary to Harbach's supposition, users' failure to report on certain risks (e.g., phishing) showed their unawareness of those risks. We find that users are aware of those threats but have already implemented coping strategies that eliminate the need to worry about those risks. Additionally, we provide greater detail regarding the harm that users associate with various risks.

2.3 Cost and Benefit Tradeoffs

Other work in usable security provides evidence that users weigh costs and benefits when deciding what security advice to adopt. For example, Fagan et al. examined the basic question of why some people follow security advice but others do not [11]. They find that the benefits of following security advice are rated higher by those who follow the advice than by those who do not. Likewise, the risks and costs of not following it are rated higher by those who follow advice. They find that individual concerns are rated higher than social concerns, confirming work by Anderson et al. [3].

Beautement et al. conducted interviews of 17 employees from two companies to determine why they do or do not comply with security policies [5]. Their findings suggest that business users weigh the cost and benefit of compliance to design which policies to adopt. Further, they theorize that users have a limited compliance budget that must be managed, restricting users' focus to the security practices that would be most effective. Our work shows that home users have an analogous 'compliance budget' that dictates which security behaviors they are willing to adopt.

Stobert and Biddle [37] conducted a grounded theory study regarding users' behaviors in managing passwords. They found that while users took steps commonly considered insecure (e.g., writing down passwords), these choices were often rational and represented a self-management of personal resources. Our results complement their results and demonstrate that this type of rationale extends beyond password behavior into all parts of a user's digital life.

Ion et al. [23] conducted two online surveys to identify discrepancies between expert and non-expert security practices in order to improve security education for non-experts. They report that non-experts focus on using anti-virus software, making strong passwords, changing passwords frequently, watching for phishing, and visiting trusted websites. Our results reveal similar practices among non-expert users and further discuss how they select these behaviors and reject others that were adopted by experts in Ion et al.'s study.

Redmiles et al. [31] investigate the acquisition of security behaviors by focusing on how users decide which items of security advice to follow and which to ignore. They find that users commonly learn about security behaviors from the media, peers, family, and IT professionals. They found that the trustworthiness of computer security advice was largely correlated with the perceived trustworthiness of the source, in contrast with physical security advice which individuals felt capable of assessing on their own. Participants described many more reasons to reject security advice than to put it into practice, including concerns about its role as a marketing tactic and the perception that the security of their data was the duty of service providers. This work contextualizes ours by characterizing the external sources from which users learn their coping strategies. Our work extends this idea by describing additional elements that factor into the equation of how users determine which security behaviors and mechanisms to adopt.

2.4 User Motivation and Understanding

Some work has also explored user motivation. Adams and Sasse [1] challenged the view that users are not motivated to behave securely by exploring why they ignore corporate password policies. They argue that a lack of user-centered design is a result of insufficient communication between designers and users.

Furnell et al. conducted a qualitative study of novice Internet users and their awareness of, attitudes toward, and experience with online security [17]. Their work concludes that "users do not seem sufficiently interested or motivated to protect themselves" and posits that developers should discontinue reliance on users or remove users' choices in matters of security. In contrast, our work finds that users do take responsibility for their online security postures and that their decisions to reject additional security behavior are rational considering the cost of adopting those behaviors and the limited harm that would be prevented.

Kang et al. applied grounded theory to explore the connection between users' understanding of the Internet and their privacy practices [25]. Their results indicate that people with a better understanding of the Internet perceive more threats, but their analysis found no connection between the level of understanding and security practices. Relatedly, Forget et al. [15] compared users' self-reported engagement with computer security against their actual security practices and found that there is not a strong link between the two. Our results suggest that this disconnect between knowledge of threats and security practices can be explained by users' unwillingness to compromise the usefulness of the Internet. Additionally, we find that even if users are aware of certain risks, they will ignore them if they have a low probability of occurring or have minimal potential harm.

3. METHODOLOGY

We conducted an IRB-approved user study to interview individuals about their perceptions and behaviors related to online risk, risk mitigation strategies, encryption, and browser security indicators. This section gives an overview of the interview process, discusses participant recruitment and demographics, and describes our methodology for analyzing the interviews. The full details are in Appendix A.

3.1 Interview Process

Interviews were performed over a five-day period beginning November 2, 2015. In total, 23 participants were interviewed. Interviews lasted between 15 and 45 minutes, with most taking roughly 25–30 minutes. Each participant was compensated $25 USD irrespective of interview duration.

Interviews were mostly conducted in either the home or place of employment of each interviewee. This was done to avoid requiring participants to meet at a specific location, as well as to make participants feel more at ease during the interview. In two cases where this was not an option, participants were instead interviewed in public locations.

At the start of the interview, the participants were presented with a consent form notifying them that the interview would be recorded. After completing the consent form, participants completed a short demographic survey.

Interviews were semi-structured. Participants were informed that the survey was not an assessment of their understanding of the Internet or its security, but rather was designed to help our research group understand what people thought of these issues so that we might build systems that addressed their concerns. Participants were encouraged to share all of their thoughts and opinions, no matter how off-topic those might seem. The interviewer took great effort to allow participants to guide the discussion, such as changing the sequence of topics or discussing topics that were not a part of the interview guide.

3.2 Interview Guide

The interviewer was provided with an interview guide containing an ordered list of questions intended to spark discussion as necessary.

First, participants were asked how they used computers and mobile devices in their day-to-day lives. This included how many devices they owned, what they used them for, and how often they were used. Participants were also specifically asked to detail the types of online activities they engaged in.

Second, participants were asked to describe the risks and threats they were concerned about when using the Internet and whether they had personally suffered harm online. They were then asked what steps they took to protect themselves while using the Internet. This portion of the interview lasted the longest.

Third, participants were queried about encryption.[1] Participants were shown a browser address bar with an HTTPS lock icon, asked whether they had previously seen the icon, and what they thought it meant (see Figure 1). They were also asked whether they had heard the term encryption before. Those indicating they recognized the term were asked what they thought it entailed. Participants were then asked what sensitive information they had previously communicated over the Internet (either through Facebook or email), whether they would like the ability to encrypt those messages in the future (e.g., encrypted email), and how they would like that process to work.

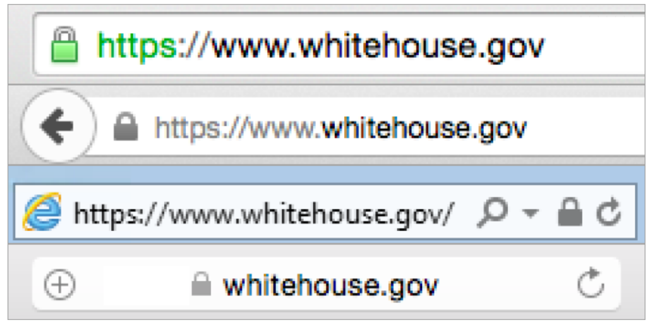


Figure 1: Examples of the lock icon from Chrome, Firefox, Internet Explorer, and Safari.

Fourth, participants were asked about their experience with security notifications. They were asked to describe what they liked and disliked about the notifications they had seen and to describe their ideal notification.

Fifth, participants were shown invalid TLS certificate warnings from the major browsers. (see Figure 8 in Appendix A.4). They were asked whether they had seen these warnings before, and if so, what they thought the warning was describing. They were then asked if they ever ignored this warning by clicking through, and if so, under what circumstances they would make this decision. They were also asked how often these warnings interfered with their day-to-day tasks, and whether they wished they would go away.

Sixth, participants were asked whether they had any other thoughts or opinions they would like to share. In this portion of the interview, participants were free to talk about whatever subjects they wished, and the interviewer avoided guiding this discussion by only asking clarifying questions as needed.

3.3 Participants

We recruited adult participants (aged 18 or older) living in Gig Harbor, Washington, U.S.A. This location was not proximal to our institution. Our aim in choosing a remote location was to seek opinions from individuals dissimilar to our research group members. Similarly, the location allowed us to survey a non-university population, which is distinct from most studies in the literature.

Initially, we tried to recruit participants from the wider Seattle/Tacoma region using craigslist, but ultimately this method yielded no participants.[2] We then posted flyers at several public locations (e.g., library, church), which resulted in recruiting eleven participants. Of these participants, one introduced us to five coworkers (teachers), six introduced us to their spouse, and one introduced us to their sister.

Demographic data for participants is shown in Table 1. In general, our participants skewed female. Participants were nearly all middle-aged and older. Most participants were currently married, with all participants having some children, the ages of whom ranged from infants to adults. Participants

[1] We asked participants about secure messaging (and by extension encryption) and TLS warnings because these are all topics that our group is actively exploring.

[2] In hindsight, it would have been possible to conduct video interviews online with several individuals that were interested, but were unable to meet in person.

		Total	%
Gender	Male	7	30%
	Female	16	70%
Age	25–34 years old	1	4%
	35–44 years old	7	30%
	45–54 years old	5	22%
	55 years or older	10	43%
Education	Some higher education, no degree	4	17%
	College or university degree	13	57%
	Graduate Education	6	25%
Career	Homemaker	7	30%
	Special Education Specialist	5	22%
	K-12 Teacher	3	13%
	IT Support	2	9%
	Medical Professional	2	9%
	Computer Scientist	1	5%
	Entrepreneur	1	4%
	University Professor	1	4%
	Unknown	1	4%
Marital Status	Married	21	91%
	Single	1	4%
	Other	1	4%
Have Children	Yes	23	100%
	No	0	0%

Table 1: Participant Demographics

had all received at least some higher education, with the majority having finished a university or graduate degree.

3.4 Limitations

Due to the nature of our methodology, our findings are subject to some limitations. First, the semi-structured interview process has some standard limitations. For example, interviewees have a desire to appear knowledgeable and competent to the interviewer, leading them to report security behaviors that exceed their actual behaviors.[3]

Second, the homogeneous nature of our interview sample's demographic—and the city from which it was drawn—limits the generality of our results. Future work could replicate this study with different populations, as well as examine specific results in a more quantitative and large-scale fashion (e.g., Mechanical Turk survey).

4. DATA ANALYSIS

After all the interviews had been completed, the audio from each interview was transcribed. These transcripts served as the primary resource used during our analysis of the data, though the audio data was referenced whenever there was ambiguity regarding the text or tone of a particular line. Throughout this paper, when quoting participants, they are labeled as P[1–23], respective to the order in which they were interviewed. This transcribed data, along with materials produced during our analysis are available at `https://soups2017.isrl.byu.edu`. Transcripts have been modified to remove personally-identifying information.

Our analysis of the data followed a four-stage grounded theory approach (open coding, axial coding, selective coding, and theory generation). Throughout the discussion process, we kept detailed research notes that outlined the thought process underlying our codes, concepts, categories, and theories. These notes were consulted frequently to guide our process. As is often the case in grounded theory, these notes were just as important—if not more—than the concepts and categories derived from the various phases of coding.

In the first stage, our research group reviewed each transcript phrase-by-phrase and word-by-word to assign codes that classified users' responses. These codes were generated using a mixture of open coding (assigning a code that summarizes the participant's statement) and in situ coding (using the participants own words as the code). To ensure that we were assigning the correct meaning to various codes, we paid attention to the context of each statement and reviewed the interview audio as needed to hear the tone the participant was using.

In the second stage, we used the constant comparative method to group codes into concepts. Specifically, we collapsed distinct codes referring to the same topic (e.g., one was an open code, the other in situ) into a single code, reducing the original set of 2,442 codes to a more manageable 503 codes.

In the third stage, we printed each code onto an index card, then organized those index cards into related categories. In total, there were nine categories describing participants' responses: The Internet, Nothing Is 100% Safe, Online Threats, Harm, Coping Strategies, Encryption, Browser-Based TLS Indicators, Secure Messaging, and Notifications. Within these groups, we drew and labeled connections between related concepts. We also drew and labeled connections between the categories.[4] Figures for each category are found throughout the paper and in the Appendix.

In the fourth and final stage, we used the categories, their connections, and our results to derive a theory describing the process users employed in selecting which security behaviors to adopt and which to reject. This theory is based both on the raw data we collected and our analysis of that data. As it is drawn from only 23 participants, it is not conclusive but does provide a theory grounded in the data we gathered.

4.1 Limitations

Due to the nature of grounded theory, our analysis of the data represents one view on that data. Different researchers coding the same data are likely to focus on different aspects leading to distinct categories, connections, and theories. We generated several theories during our research. This paper focuses only on what we determined to be the strongest and most compelling theory. To address this limitation, we will make the transcripts of our interview public.

5. THEORY

The result of our analysis was the generation of a theory that describes the process by which users select their online security posture. Before discussing the process, it is important

[3]Interestingly, participants in our studies often freely admitted that they were doing less than they should. While it is likely that illusory superiority had an effect, it is also possible that the snowball sampling led users to feel the interviews were more personable (i.e., recommended by their friends). This most likely led to more honest answers.

[4]Due to the visual complexity of the complete theory graph, we have not included it in this paper.

to understand the context (i.e., environment) under which this process operates. This context is dominated by two components, the utility of the Internet in users' lives and that users believe perfect security is not achievable.

The participants in our study unanimously indicated that the Internet has been transformational in their lives (see Figure 2). All participants derive value from their use of the Internet, with many noting that it was now a central part of their lives. For example, P13 emphatically expressed, *"I love the Internet. It's become my world."* For others, the Internet has allowed them the freedom to live their lives as they want. P3 described this saying,

> *[The Internet] made our whole home schooling process possible. So our kids grew up to be different than they would have been if they had just gone to the local public school, which was real poor quality. [...] If we had been teaching our kids ten years sooner, it would have just been a huge impact. I mean our lifestyle would not have been possible before the Internet.*

Participants also indicated near unanimously that no matter how much effort was put into strengthening their online security posture, it was impossible to be 100% safe (see Figure 3)—as described by P19, *"I don't think there's ever a place that is perfectly safe."* This viewpoint was derived from three sources:

1. Dramatized depictions of hackers on television and in movies, where security is broken in dramatic fashion, e.g., in 30 seconds or less.

2. Frequent news reports that even companies with large security budgets were routinely compromised—P4 explained, *"there are some big companies that get hacked, that I would expect would have good security in place, but they still get hacked."*

3. An interpretation of the cyber-world as seen through the lens of the physical-world. Specifically that, like the physical world, nothing was ever completely safe—

   > *They got into [the] Pyramids; they got into King Tut's tomb. They can walk in here [at] any time, even with the doors locked. So, I guess I've come to believe there's a segment of society that's gonna make trouble for the rest of us, no matter what generation or what age or what media or [by] what means.* (P14)

Because they did not believe it was possible to completely stop an attacker, participants' security behaviors derived from a focus on addressing the most common threats and making themselves a less appealing target. As expressed by P14, *"you throw enough stumbling blocks in [an attacker's] way, they're gonna look for somebody else that's easier to take care of, to get into. I would imagine it's very much the same way with Internet and security or through encryption."* All the while, participants were careful that their security posture did not unduly affect their ability to derive value from their Internet use.

Within this context, we identified a four-step process which guides a user's selection of which security behaviors to implement. More specifically, users weigh the advantages of online activity against the cost of implementing security practices or mechanisms intended to minimize the likelihood or negative consequences of online risks. While this process is described linearly, users are fluid in their application of it. As they learn and evaluate new information, they may skip some steps or re-evaluate others.

1. **A user learns about a new security threat.** This happens by word of mouth, news reports, television shows, and movies.

2. **The user evaluates the risk presented by the threat.** If the attack seems sufficiently unlikely, they will generally ignore it.

3. **The user estimates the impact of a successful attack.** The amount of damage is congruent to the effort they are willing to expend to address the threat.

4. **The user selects an appropriate coping strategy.** This selection is based on trade-offs between the cost of implementing the coping strategy (e.g., diminished ability to use the Internet) and its ability to mitigate risk by reducing attack surface and/or impact.

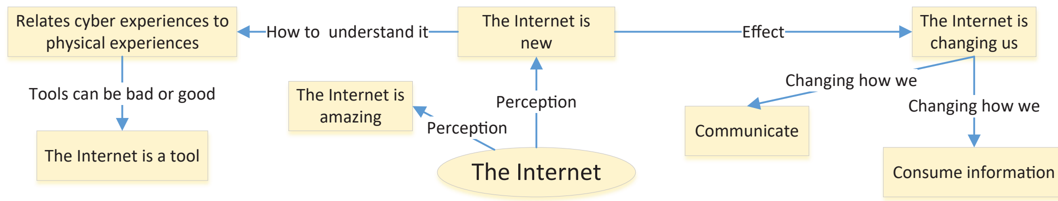
5.1 Learning about Threats

Participants reported learning about threats through four primary sources, all media-based: advertisements, news reports, television dramas, and movies. For example, roughly a quarter of participants asked the study coordinator about LifeLock, an identity protection product, noting that they had heard about it on a radio advertisement. Similarly, P11 described how she learned about cybersecurity from the nightly news:

> **P11:** *Yeah. It wasn't until a couple of weeks ago that they talked about the dark side of the Internet. I didn't know there was one until they started it. You know, it is pretty interesting.*
> **Interviewer** *Where did you hear about that?*
> **P11:** *On the news!*
> **Interviewer** *Local news?*
> **P11:** *Yes. Local news was talking about the darker side of the Internet.*

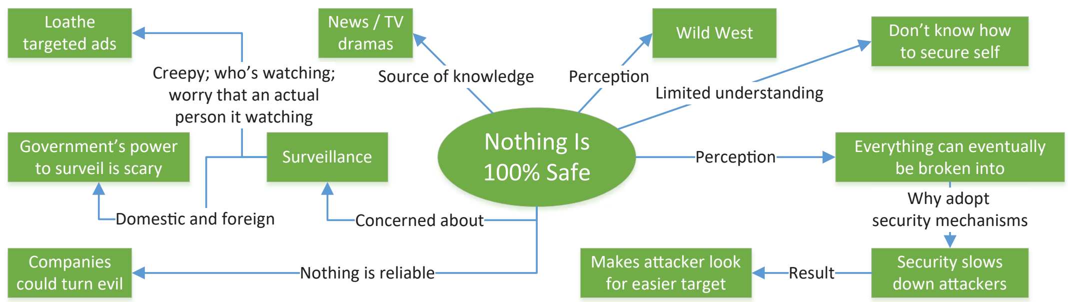
Participants' understanding of encryption and hackers was tied to television dramas or movies. When asked how strong they thought encryption was, P12 replied, *"I would think it might be fairly easy. At least from the movies, they make it sound like they try all these combinations on a computer, and then in thirty seconds, the code's cracked."*

While participants did not report learning about *new* online threats from friends, they did describe using them as a clarifying source that provided greater details regarding the threat and helped identify potential coping strategies. In several cases, participants noted that they relied entirely on their spouse as the de facto security expert. For example, P10 described her key method for evaluating the risk of unknown content: *"If I get something that I don't know, I'm not calling someone until—I actually just call my husband first."*



Participants overwhelmingly expressed wonder at all the Internet has allowed them to do. Many noted that it is a tool that can be used positively or negatively. Others pointed out that it is changing human behavior, including how we communicate and consume information.

Figure 2: The Internet Category Graph



Participants indicated that nothing could be perfectly safe. Critically, participants believed that given sufficient time, hackers could break any system; at best, security slows attackers down, causing them to choose different targets. Also, participants noted that currently trustworthy organizations—corporate and government—could become malicious in the future.

Figure 3: Nothing Is 100% Safe Category Graph

5.2 Evaluating Risk

When learning about a new threat, participants attempted to evaluate its risk (see Figure 4). Several participants noted that most threats did not imply personal risk because the chance they would be targeted was small. P22 explained that *"there's so many of us. I think that kinda helps us, too. There's so much information out there that it's highly unlikely that you'd be targeted, but you can be."* Still, P23 noted that this protection was not airtight: *"my only protection is that I am only one of 300 million. But you know, I got [...]* a year ago—that's a 1 in 10,000 chance [...]. Somebody gets picked."*[5]

The threats that participants deemed most risky (i.e., likely to affect them) were largely threats that they had previously encountered—malware, phishing attacks, inappropriate content—or which they had heard discussed frequently on the news—data permanence and surveillance. While the former category of attacks has been discussed at length in the literature (e.g., [5, 41, 20, 36]), the latter is largely unexplored.

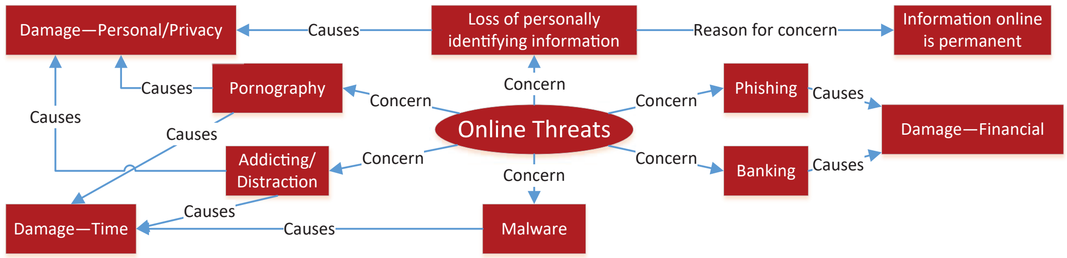
The permanence of online data without consent was a strong concern for many participants. Participants noted that once something was said or done on the Internet, it would remain forever (especially on social media). Several parents and teachers in our study indicated that they make an effort to educate children about the risks of posting information online. This threat troubled participants because once they uploaded any data they were unable to ensure it would be maintained according to their wishes. This led to a tension between using the Internet freely and ensuring that their personal data would not be used inappropriately.

Within this vein, P3 made a compelling argument that children's inability to erase past online interactions could have a chilling effect on their ability to mature:

> **P3:** *[...] there is some concern with kids using Facebook and having a personality develop online. It would be nice to somehow have an opportunity to erase that as they get older. I don't know what it will be like for this generation. We didn't—we were able to grow and mature and change, and leave behind our old selves at some point. It would be nice if there was some way that kids could—*
> **Interviewer:** *That they don't have to be haunted by the silly things they said as an adolescent.*
> **P3:** *Yeah, that things are permanent once they are leaked into the online world. That (encryption) would be very useful. I think I would feel more able to develop in classes, like writing classes, where you're submitting things and opinions. You'd feel more free to develop in that way if you knew that they weren't going to be a permanent part of your record to everybody for now and ever.*

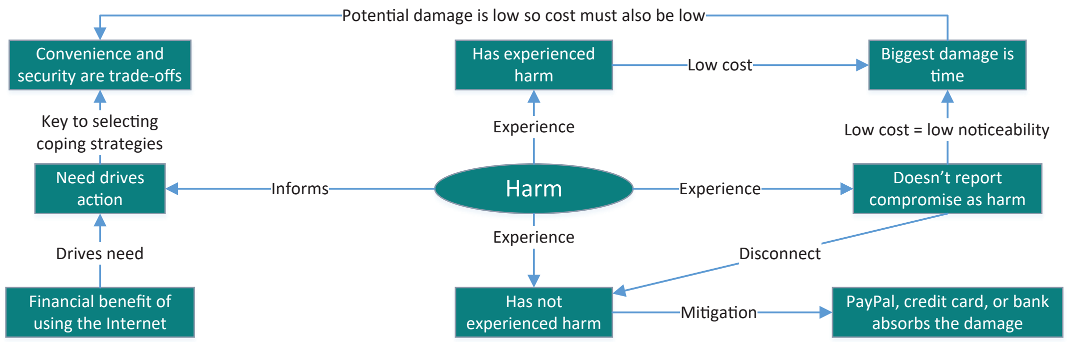
Participants also noted that surveillance of their online activities was a foregone conclusion, especially in relationship to government surveillance. P16 expressed,

[5] P23 described contracting a rare illness here, which has been redacted to preserve anonymity. The point being made is that even low-likelihood events affect *someone*.



Participants were concerned with a wide range of online threats. These threats were associated with three types of damage—financial, personal information and privacy, and time. Interestingly, malware was considered to cause no harm other than the time it took to remove it.

Figure 4: Online Threats Category Graph



Many participants had personally experienced harm or knew someone that had. Still, these experiences were not impactful because they produced no lasting consequences. This lack of impact led users to avoid strengthening their security posture.

Figure 5: Harm Category Graph

> *Well, there's rumors that [the] government watches over everything that we do and that certain words, even in your conversation on the phone, could be flag words. Then you could suddenly have a person at your door. And I don't know [...] [i]f that's a lot of conspiracy theory, or how much reality that is. But it's a possibility, because everything that's good can be used for evil, you know?*

Concerns regarding surveillance were not limited to the government, but also included companies tracking online activities: *"there's just the generalized concern about what can people see me do? How many people are watching me? Who's watching me?"* (P19). This unease was reinforced when their actions on one site would result in related advertisements being shown on an entirely unrelated site. P7 shared,

> *I hate the ads. I hate the ads! [...] [S]ay I go online and I'm looking at a certain style of shoe. And then I come back a week later, and I just open my web browser to my home page, little ads are streaming about what I was looking at.*

5.3 Estimating Impact

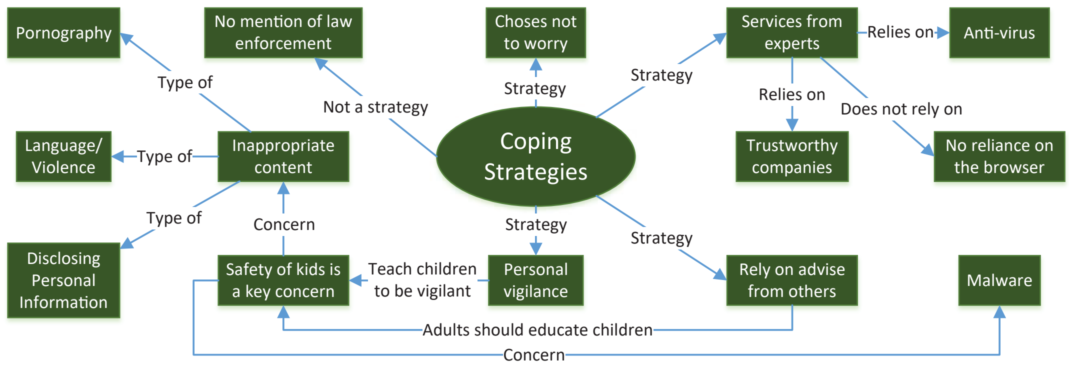
To further characterize threats, users estimated the impact threats could have on their lives (see Figure 5). Unsurprisingly, attacks which led to financial damage were viewed as the most impactful. P4 expressed that such harm was quite scary—*"something that you worked so hard for—your money, and your well-being—and then to have it disappear in a second, is a bit scary."*

While financial damage was a concern, some participants explained that it was unlikely that it would be permanent. They noted that for online shopping they used PayPal or a credit card, both of which would absorb the cost of any successful attack. As described by P3, *"We did have a credit card [company] call us, and let us know that there was a charge on there, and wondering if it was ours, and it wasn't, so they declined the charge before it even went through."* These protections reduce the damage of such attacks to only the time and effort it takes to get charges reversed.

Other than financial harm, participants expressed concern regarding their children's online safety. They worried that children had insufficient experience to avoid malware and that they were likely to disclose personal information without fully considering the ramifications of that disclosure. Additionally, they worried that it was easy for children to be exposed to inappropriate content online (e.g., violence, language)—P12 said *"my boy loves to get on YouTube and listen to Vines and stuff, and the language can be atrocious."*

For most other threats, potential harm was usually seen as negligible—usually only representing a small cost in time or effort to resolve. P14 described this issue,

> *Well, first of all, all [of a] sudden your computer is doing something that you can't get rid of, just, you know, no matter what you do, it's still there. So, the only*



Participants' key coping strategy was personal vigilance, as they believed that their security was ultimately their responsibility. Participants' primary concern regarding their children was teaching them to be vigilant online—to avoid downloading software from suspicious websites (i.e., malware), viewing inappropriate content on YouTube, and disclosing personal information on social media.

Figure 6: Coping Strategies Category Graph

way to get rid of it is to completely go back, either in time, or go through and find out which program it is and determine a way to dump it. But, that's not always easy to do.
...
It's just time-consuming! And you're looking to computers to do the opposite: allowing you more time. Instead, what you find is sometimes it takes a lot more time to either solve a problem or... If something goes down, then it becomes an issue.

Surprisingly, due to the perception that costs in time and effort are negligible, participants who had been successfully attacked did not always relate those attacks with harm. When initially asked if they had experienced harm online, they would report that they had not. Later in the interview when discussing other topics, it would become clear that they had previously been compromised (e.g., installed malware, stolen credit card). When asked about this discrepancy, participants indicated that because it had been so easy to resolve, they did not consider it harm. This attitude towards harm helps explain why users are quick to ignore threats that only result in minor loss of time or effort.

5.4 Selecting Coping Strategies

Users select coping strategies based on their evaluation of trade-offs—harm addressed vs. cost to implement (see Figure 6). As the Internet is a critical piece of users' lives, even minor reductions in its utility can be viewed as costly. Also, as participants did not believe they could be perfectly safe, their effort was focused on the most effective coping strategies for reducing their likelihood of being attacked and/or minimized the negative consequences of victimization. Non-selected strategies were largely rejected because they were viewed as having marginal value that did not outweigh their cost of adoption. Depending on how users weigh these various factors, the selected coping strategies can be wildly different.

On one extreme, P13 stated that she did nothing to protect her online security:

Interviewer: *When you are using the Internet, are there any risks or threats that you worry about?*
P13: *No. I bank online and I don't care. I know people who worry about that, I don't.*
Interviewer: *So you never have any concern, regardless of what you are doing on the Internet.*
P13: *That's true. I should have concerns. I know I should.*
Interviewer: *Tell me why you don't have any concerns.*
P13: *Because I don't want to.*
Interviewer: *Can you elaborate on that?*
P13: *OK. I don't want to get trapped. I want to use the convenience of the Internet, and not feel scared of the Internet. People think that their identity is going to get stolen, and it can be, or their bank account is going to be gotten into. Which I understand it can be. Don't care. Because I don't want it trap me emotionally.*

While this attitude, at first, seems quite flippant, upon further examination, it becomes understandable. This participant was an entrepreneur whose business relied on access to plans stored in county buildings. Originally, access was physical, requiring transit to the buildings in question, a slow and time-consuming process. More recently, these documents had been uploaded to the Internet, saving her time and increasing the profitability of her business several times. To her, the Internet represents her livelihood, and any reduction in its utility represents a loss of money to her. In comparison, the coping strategies do little to protect her—the time she saves by using the Internet far outweighs the time she would lose by fixing her computer or working with a credit card company to roll back a transaction. As such, she is making a cost-efficient and rational choice.

On the other extreme, several participants were unwilling to perform any financial transactions (e.g., banking, shopping) online. In their eyes, third-parties (e.g., PayPal, credit cards) could not do enough to protect their financial safety, and the potential harm of financial compromise far outweighed any convenience brought by the Internet. As described by P11,

P11: *I feel very uncomfortable doing banking online. I always have.*
Interviewer: *I would love to hear why.*
P11: *You know, I was doing pretty good with the con-*

cept, and then [...] one of the banks... Bank of America got tapped... somebody got tapped into, and I thought 'Oh, if it's that easy, I just, no, there is somebody else who is spending more time out there then what is needed.' I can go. I'm social; I can go and say hello and get my banking done.

In general, participants fell between these two extremes, selecting to implement coping strategies which had acceptable trade-offs. The coping strategy described as most important was that of personal vigilance—namely, being careful about what sites they visited, what links they clicked, and what files they downloaded. This strategy was selected because it had low cost—it is easy-to-implement and easy-to-bypass as needed—and also because participants felt that they needed to take *personal* responsibility for their online safety. For example, P1 expressed,

> *That, of course, reminds you, that I myself am responsible for monitoring my personal information. Especially as it regards credit and banking, those kind of things. It is up to me to monitor those things on a consistent and regular basis.*

In addition to personal vigilance, other commonly reported coping strategies included installing an anti-virus, setting up a web filter (e.g., OpenDNS, NetNanny) to block inappropriate content, using PayPal and a credit card for online shopping, and relying on services provided by large, credible companies. As these are well studied coping strategies, we do not discuss them in further detail. Conspicuously, two coping strategies were absent from participants' responses—law enforcement and the browser. While participants did mention browser-based TLS securing indicators when asked about them, they did not proactively report these types of features when asked about their online security behaviors. Law enforcement, by comparison, was not mentioned even in passing, even when it came to descriptions of financial risks or sources for advice regarding online safety.

Interestingly, after describing what security strategies they had adopted, several participants indicated that they chose not to worry about remaining threats. The reasoning behind this was the remaining threats were less likely to impact their lives, that they were unaware how these threats could be addressed, and that they didn't want to worry while using the Internet (similar to the sentiment expressed by P13). For example, P4 stated, *"Well, there is a reason to worry, but I don't know what to do about it, so I can't obsess about it, get all panicky. Cause I don't know what to do.*

Ultimately, regardless of their selected security, participants were acting rationally based on the context of their Internet usage and their understanding of threats, potential harm, and trade-offs for various coping strategies. In each case, users were able to give a cogent explanation for why they adopted some coping strategies while rejecting others. Our results suggest that it is counterproductive to either browbeat users into compliance or to bypass them entirely. Instead, if security tools can be better aligned with users' environments and needs, then adoption is much more likely.

6. ADDITIONAL TOPICS

In addition to the topics covered in our theory, participants reported interesting thoughts regarding several additional security topics—encryption, browser-based TLS security indicators, and secure messaging. The category graphs for these remaining topics are in Appendix B.

6.1 Encryption

Two-thirds of participants had an understanding of the basic principles of symmetric encryption, that it *"keep[s] others from being able to see things they shouldn't."*[6] Participants referred to the process of encryption as "scrambling" data, and half were aware that it involved a shared secret.[7]

In accordance with participants' belief that nothing is 100% safe, participants did not believe that encryption is impenetrable, noting that a determined attacker could either find a way around the encryption or a way to break the encryption. Participants indicated that it would take *"huge, huge computers with lots of processing power"* to break an encrypted message. They also described breaking encryption as necessitating *"savvy"* reasoning and that while it might not keep everyone out, it would take skill that *"probably 95% of the population doesn't have."*

While several participants used encryption tools as part of their job, none used them in their personal lives. When asked if they could identify any personal uses (i.e., non-business, non-HTTPS) for encryption, almost half of participants indicated that they did not see a use for it, either because they did not upload sensitive information online or because they doubted that encryption of online information could ever be sufficiently secure (see Section 6.3).

Of the participants that identified uses for encryption in their personal lives, they mentioned protecting financial data, cloud data, work documents, and day-to-day communications. For several participants, encryption intrigued them because it offered a potential solution to two threats that they lacked adequate coping measures for: government surveillance and data permanence (see Section 5.2). For example, P3 indicated that encryption could be a solution to children's online interactions being too permanent (see Section 5.2).

6.2 Browser-based TLS Security Indicators

There are a wide range of papers that examine the effectiveness of TLS warnings in browsers [10, 38, 2, 13]. Collectively they find that many users ignore TLS warnings, but that over time users have become more likely to heed these warnings. We questioned participants regarding the browser's TLS indicators—HTTPS lock icon and TLS warning page—to better understand what they thought these indicators meant and why they sometimes choose to ignore them.

When presented with images of these indicators, it quickly became apparent that participant's mental models largely failed to account for the existence of connection-level attacks. Instead, participants associated the TLS indicators with site-level safety.[8] Importantly, we found that these misconceptions were directly correlated with insecure behavior.

[6]Other than one participant who was a software developer, participants never mentioned public key cryptography.
[7]The terms used to described the shared secret included calling it a key (most participants), a password, a credential, or a code.
[8]Similar misconceptions about connection-level security have also been observed in more technical users [12].

6.2.1 HTTPS Lock Icon

Participants largely felt that the lock icon indicated that the site was "safe" place to do business. For example, P9 said,

> *Well, to me it means that it is a secure site. That other people are not just going to be able to get into what I have put in. I'm sure that there are ways to do that, but they've made it harder—hopefully. So, that's what it has meant to me.*

Others thought that the lock icon indicated that the website was "locked", and that it would require credentials to access. For example, P8 said the following about the lock icon:

> *That would make me think that I need a password to get it. A password or login to get in. That it's secure. That it's only for those people, where you have to create an account, or for those that have already created an account.*

Regardless, participants indicated that they did use the lock icon to determine which sites they should use. P9 explained, *"if I'm about to use my credit card, I do look for it to make sure it is there. Sometimes if it is not there, I won't purchase. I'll just say, 'Well, I can go find it at the store.'"* Other noted that they used the lock icon to ensure they were on the "real" website. P15 stated, *"like Bank of America, if it's locked, they are telling me it is their website, and that it is the right website."*

While at first glance, users' attention to the lock icon might seem like a positive sign, it has troubling implications. Phishers could take advantage of users trust in the lock icon by transmitting their phishing websites over TLS, leading users to be more likely believe that the website is legitimate and safe. This idea is lent credence in the next subsection where we discuss how a similar misconception causes participants to click through TLS warnings. As such, efforts to increase user attention towards the lock icon [12] may end up being counterproductive.

6.2.2 TLS Warning Page

In contrast with the shared, albeit vague, understanding of the browser lock icon, the browser's TLS warning pages were nearly unilaterally met with confusion, though several noted that it meant that *"I'm in trouble"* (P21). Many participants expressed confusion when seeing the warning—P7 said, (*"I don't know what a security certificate is. I've seen [the warning before], but I have no idea what that is."* Still, others thought it was an indication of the site's trustworthiness, similar to their HTTPS lock icon misconceptions.

Overall, participants reported seeing these warnings rarely—at most once a month. Most often these warnings were seen when accessing the participant's employer's intranet, which was described by some users as being rife with sites that required clicking through the TLS warning.

Most participants reacted to these warnings by opting to back off entirely, particularly if they felt at all uncertain. Others indicated that they would ignore the warnings only if they were consuming information and not inputting information. For example, P3 indicated that *"if I'm just looking for information, I have just ignored that. But if I am thinking of shopping, I think I have thought, 'I'm not going here.'"*

Disconcertingly, some participants believed that this warning was a judgment of the trustworthiness of the website being visited. For example, P12 (a Chrome user) said the warning meant that *"if there is an untrusted site that Google doesn't quite know, they are saying 'We don't really know about these guys, and if you want to continue, you can, but we don't really know about them.'"* This misconception led participants to believe that they could safely ignore the warning if it were for a website that they "knew" was safe. P7 stated, *"Well, if I see it, and I am going into some place I have never been before, then I will probably just not go. If it is a place that I know is OK, because I have been there before, then I usually go ahead."*

In these situations, participants attributed the error to a misconfiguration by the browser or website. P19 suggested, *"well, maybe they've just done an update or something like that and there's a glitch in the update."* Alarmingly, the choice to bypass the TLS warning was often associated with high-value sites (e.g., Amazon, email)—these sites were well known to the user—in direct contrast to the fact that the TLS warning is most likely to indicate an attack when it appears for these sites.

6.3 Secure Email and Messaging

Most participants indicated that they had no need for secure messaging and secure email in particular. Many noted that they rarely needed to send sensitive documents, and when they did (e.g., loan application) the company would request that those documents were uploaded directly to the company through a web portal. When asked how they transmitted sensitive data person-to-person, participants indicated that they would share it in person, over the phone, or through text. They viewed these activities as more secure because they felt that each of these transactions was ephemeral—requiring an adversary to actively be targeting them, while online communication (e.g., email) was permanent. For example, P12 described this at length:

> **Interviewer:** *And do you think phones are more secure than email?*
> **P12:** *I think they are, in the standpoint that there would have to be someone bugging your phone and catching it immediately. Whereas if you send it on the Internet, or email, it's logged, and anybody can... So, I guess just your window of opportunity is a lot larger in an email, or on a search engine. Whereas a phone, they would have to be listening that plus or minus maybe five seconds to get the information that they need. To my knowledge, I don't know that anyone is recording, for long periods of time, my phone conversations.*
> **Interviewer:** *So it is really... it seems to be that permanence.*
> **P12:** *Permanence.*
> **Interviewer:** *That with email if you send it once, they can come back later.*
> **P12:** *Exactly! Your window is much larger that you leave yourself exposed. Whereas a phone call you have only got five, six seconds. Blah blah blah blah blah, there is the credit card number. That's the biggest reason. [...] Everybody says that things are so permanent on the Internet. They dig up stuff that is twenty years old. Then they find dirt and information on politicians, and stuff like that. They're finding emails from Hillary Clinton, from years past, when she was at the State Department.*

> *They are digging up all sorts of stuff. And now they are saying SnapChat—where it is supposed to snap a picture and be gone instantly—ummm... is not.*

In most interviews, we asked participants to imagine a hypothetical secure email system that was both usable and fully secure. We then asked participants if they would find a use for such a system, and if so how would they use it. While most participants expressed an interest in such a secure email system, many of those who were interested struggled to identify when they could use it—they only rarely had the need to send sensitive information. Most were interested in its ability to make email messages containing sensitive information ephemeral, self-destroying after the information was no longer needed.[9] The two medical professionals were especially interested in the possibility that secure email could substantially expedite the process of sharing medical information between institutions.

Interestingly, when asked to imagine the hypothetical secure email system, several participants pushed back and stated that they *"would be very skeptical that something like that would ever exist"* (P12). This attitude was tied to their perception that nothing was 100% safe and therefore no secure email tool could protect their sensitive information from determined parties.

7. CONCLUSION AND FUTURE WORK

Our interviews demonstrate that users' online security posture is guided by an analysis of the cost-benefit trade-offs of various coping strategies, informed by their understanding of risks, potential harm, and the context of their online activities. While a user's set of coping strategies is insufficient to address all potential threats, those strategies are usually sufficient to protect them from the harm that they are most concerned about. While these results are drawn from a limited sample of participants and do not necessarily generalize to the entire population, they still provide a helpful guide for what future research can explore.

Because users make rational decisions and are actively engaged in considering their personal security posture, it means that they can be influenced to improve their security posture. While there are many areas where research could be done to better address user needs, we discuss below five areas that stood out as important and achievable as we interviewed participants and analyzed their responses.

Security Recommendations. Participants prefer coping mechanisms that have the greatest impact on reducing their attack surface—i.e., they are not interested in security behaviors that have marginal gains. For example, Florencio et al. [14] show that passwords that are 8–10 characters long are generally resistant to online attacks, whereas passwords of length 18–20 are needed to resist offline attack—passwords of length 11–17 offer marginal security gains at significant cost to users. By recommending that users select password 8–10 characters long, the users can focus on a coping strategy that has a significant impact, without trying to guilt them into adopting longer passwords that have either marginal benefit or become overly-difficult to remember. Future research should follow this cue for passwords, and distinguish which recommendations have low cost and high impact, and which only offer marginal returns.

User Education. Our study showed that most participants learned about online security through media—i.e., news reports, television shows, and movies. Ideally, the community could influence these mediums to correctly portray cybersecurity issues, but this is unlikely. Alternatively, participants noted that they and their children regularly watch content on YouTube and similar services. This presents a compelling avenue for disseminating accurate cybersecurity information to the masses. Future research could explore how to structure such online videos to both educate and to attain maximum dissemination. Based on several participants' responses, a good place to start would be Whiteboard-style videos.

Privacy-preserving Systems for Children. The literature on strong privacy-preserving systems is primarily focused on high-security situations—e.g., political dissidents. The resulting security model is often very strict and leads to relatively low usability. According to their parents (the participants), children are often unaware of the potential harm of disclosing personal information online and are thus unlikely to pay the high usability cost of adopting such solutions. Future research should examine how privacy-preserving technologies can be better adapted to the needs of children for use as they grow up.

Browser Indicators. Users are primarily concerned with the safety of the sites they are visiting, while browsers display information regarding the security of connections. While connection security is an important metric, it does not fully address users' primary concern. Future research should explore how the browser can be used to inform users regarding the safety of the sites they visit. This could have more impact than focusing on making users pay more attention to an indicator (i.e., HTTPS lock icon) that they misunderstand.

Secure Email. Participants indicated that their greatest worry regarding email was its permanence, yet current secure email research is focused on usability [43, 35, 34], not message permanence.[10] Future research should explore how to make email more ephemeral so that users can control the permanence of their messages.

8. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. CNS-1528022 and research sponsored by the Department of Homeland Security (DHS) Science and Technology Directorate, Cyber Security Division (DHS S&T/CSD) via contract number HHSP233201600046C. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the Department of Homeland Security. Also, this work was supported by Sandia National Laboratories, a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

[9] We note neither PGP, S/MIME, nor current research into secure end-to-end email encryption address this need.

[10] While short-lived keys address this problem, current systems do not actively support this use case.

9. REFERENCES

[1] A. Adams and M. A. Sasse. Users are not the enemy. *Communications of the ACM*, 42(12):40–46, Dec. 1999.

[2] D. Akhawe and A. P. Felt. Alice in warningland: A large-scale field study of browser security warning effectiveness. In *22nd USENIX Security Symposium*, pages 257–272, 2013.

[3] C. L. Anderson and R. Agarwal. Practicing safe computing: a multimedia empirical examination of home computer user security behavioral intentions. *Mis Quarterly*, 34(3):613–643, 2010.

[4] C. M. Angst and R. Agarwal. Adoption of electronic health records in the presence of privacy concerns: The elaboration likelihood model and individual persuasion. *MIS quarterly*, 33(2):339–370, 2009.

[5] A. Beautement, M. A. Sasse, and M. Wonham. The compliance budget: Managing security behaviour in organisations. In *Proceedings of the 2008 New Security Paradigms Workshop*, pages 47–58. ACM, 2009.

[6] J. Bonneau, C. Herley, P. C. Van Oorschot, and F. Stajano. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *2012 IEEE Symposium on Security and Privacy (SP)*, pages 553–567. IEEE, 2012.

[7] S. Buckley. Hackers stole 21.5 million Social Security numbers in government breach. `https://www.engadget.com/2015/07/09/hackers-stole-21-5-million-social-security-numbers/`, July 2015. Online; accessed 21-September-2016.

[8] J. Clark and P. C. van Oorschot. SoK: SSL and HTTPS: Revisiting past challenges and evaluating certificate trust model enhancements. In *2013 IEEE Symposium on Security and Privacy (SP)*, pages 511–525. IEEE, 2013.

[9] J. Corbin and A. Strauss. Grounded theory research: Procedures, canons and evaluative criteria. *Zeitschrift fÃijr Soziologie*, 19(6):418–427, 1990.

[10] S. Egelman, L. F. Cranor, and J. 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. ACM, 2008.

[11] M. Fagan and M. M. H. Khan. Why do they do what they do?: A study of what motivates users to (not) follow computer security advice. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 59–75. USENIX Association, 2016.

[12] A. P. Felt, R. W. Reeder, A. Ainslie, H. Harris, M. Walker, C. Thompson, M. E. Acer, E. Morant, and S. Consolvo. Rethinking connection security indicators. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, 2016.

[13] A. P. Felt, R. W. Reeder, H. Almuhimedi, and S. Consolvo. Experimenting at scale with Google Chrome's SSL warning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 2667–2670, New York, NY, USA, 2014. ACM.

[14] D. Florêncio, C. Herley, and P. C. Van Oorschot. An administrator's guide to internet password research. In *LISA*, pages 35–52, 2014.

[15] A. Forget, S. Pearman, J. Thomas, A. Acquisti, N. Christin, L. F. Cranor, S. Egelman, M. Harbach, and R. 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, Denver, CO, June 2016. USENIX Association.

[16] L. Franceschi-Bicchierai. Hacker tries to sell 427 million stolen MySpace passwords for $2,800. `http://motherboard.vice.com/read/427-million-myspace-passwords-emails-data-breach/`, May 2016. Online; accessed 21-September-2016.

[17] S. Furnell, V. Tsaganidi, and A. Phippen. Security beliefs and barriers for novice internet users. *Computers & Security*, 27(7-8):235 – 240, 2008.

[18] R. H. Gass and J. S. Seiter. *Persuasion: Social influence and compliance gaining*. Routledge, 2015.

[19] GeekTime. Millions of victims lost $12.7b last year falling for Nigerian scams, 2014.

[20] M. Harbach, S. Fahl, and M. Smith. Who's afraid of which bad wolf? A survey of IT security risk awareness. In *2014 IEEE 27th Computer Security Foundations Symposium (CSF)*, pages 97–110. IEEE, 2014.

[21] T. Herath and H. R. Rao. Protection motivation and deterrence: a framework for security policy compliance in organisations. *European Journal of Information Systems*, 18(2):106–125, 2009.

[22] C. Herley. So long, and no thanks for the externalities: The rational rejection of security advice by users. In *Proceedings of the 2009 New Security Paradigms Workshop*, NSPW '09, pages 133–144, New York, NY, USA, 2009. ACM.

[23] I. Ion, R. Reeder, and S. Consolvo. "...no one can hack my mind": Comparing expert and non-expert security practices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 327–346, Ottawa, July 2015. USENIX Association.

[24] M. L. Jordan Smith. Not so securus: Massive hack of 70 million prisoner phone calls indicates violations of attorney-client privilege. `https://theintercept.com/2015/11/11/securus-hack-prison-phone-company-exposes-thousands-of-calls-lawyers-and-clients/`, November 2015. Online; accessed 21-September-2016.

[25] R. Kang, L. Dabbish, N. Fruchter, and S. Kiesler. "my data just goes everywhere:" User mental models of the Internet and implications for privacy and security. In *Eleventh Symposium on Usable Privacy and Security (SOUPS 2015)*, pages 39–52, Ottawa, July 2015. USENIX Association.

[26] K. Kaspersky. Report: Ransomware in 2014–2016, 2016.

[27] R. LaRose, N. J. Rifon, and R. Enbody. Promoting personal responsibility for internet safety. *Communications of the ACM*, 51(3):71–76, 2008.

[28] D. Lee, R. Larose, and N. Rifon. Keeping our network safe: a model of online protection behaviour. *Behaviour & Information Technology*, 27(5):445–454, 2008.

[29] C. Osborne. 'discreet' cheating website Ashley Madison suffers data breach. `http://www.zdnet.com/article/discreet-cheating-website-ashley-madison-suffers-data-breach/`, July 2015. Online; accessed 21-September-2016.

[30] R. E. Petty and J. T. Cacioppo. The elaboration likelihood model of persuasion. In *Communication and persuasion*, pages 1–24. Springer, 1986.
[31] E. M. Redmiles, A. R. Malone, and M. 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.
[32] C. Riley. Insurance giant Anthem hit by massive data breach. `http://money.cnn.com/2015/02/04/technology/anthem-insurance-hack-data-security/`, February 2015. Online; accessed 21-September-2016.
[33] R. W. Rogers. A protection motivation theory of fear appeals and attitude change1. *The journal of psychology*, 91(1):93–114, 1975.
[34] S. Ruoti, J. Andersen, S. Heidbrink, M. O'Neill, E. Vaziripour, J. Wu, D. Zappala, and K. Seamons. "we're on the same page": A usability study of secure email using pairs of novice users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 4298–4308, New York, NY, USA, 2016. ACM.
[35] S. Ruoti, N. Kim, B. Burgon, T. Van Der Horst, and K. Seamons. Confused johnny: when automatic encryption leads to confusion and mistakes. In *Ninth Symposium on Usable Privacy and Security (SOUPS 2013)*, page 5. ACM, 2013.
[36] R. Shay, I. Ion, R. W. Reeder, and S. Consolvo. My religious aunt asked why I was trying to sell her Viagra: Experiences with account hijacking. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, pages 2657–2666. ACM, 2014.
[37] E. Stobert and R. Biddle. The password life cycle: User behaviour in managing passwords. In *Tenth Symposium on Usable Privacy and Security (SOUPS 2014)*, pages 243–255, 2014.
[38] J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L. F. Cranor. Crying wolf: An empirical study of SSL warning effectiveness. In *8th USENIX Security Symposium*, pages 399–416, 2009.
[39] N. Unger, S. Dechand, J. Bonneau, S. Fahl, H. Perl, I. Goldberg, and M. Smith. SoK: Secure messaging. In *2015 IEEE Symposium on Security and Privacy (SP)*, pages 232–249. IEEE, 2015.
[40] L. Vaas. 154 million voter records exposed, including gun ownership, Facebook profiles and more. `https://nakedsecurity.sophos.com/2016/06/23/154-million-voter-records-exposed-including-gun-ownership-facebook-profiles-and-more/`, June 2016. Online; accessed 21-September-2016.
[41] R. Wash. Folk models of home computer security. In *Sixth Symposium on Usable Privacy and Security (SOUPS 2010)*, page 11. ACM, 2010.
[42] R. Wash and E. Rader. Too much knowledge? security beliefs and protective behaviors among united states internet users. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 309–325. USENIX Association, 2015.
[43] A. Whitten and J. Tygar. Why Johnny can't encrypt: A usability evaluation of PGP 5.0. In *Eighth USENIX Security Symposium (USENIX Security 1999)*, pages 14–28, Washington, D.C., 1999. USENIX Association.
[44] K. Witte. Fear control and danger control: A test of the extended parallel process model (eppm). *Communications Monographs*, 61(2):113–134, 1994.
[45] S.-C. Yang, W.-C. Hung, K. Sung, and C.-K. Farn. Investigating initial trust toward e-tailers from the elaboration likelihood model perspective. *Psychology & Marketing*, 23(5):429–445, 2006.

APPENDIX

A. STUDY MATERIALS

This appendix lists all the materials used to conduct the study. Personally identifying information has been replaced by bracketed text describing the relevant information.

A.1 Consent to be a Research Subject

Introduction

This research study is being conducted by [study coordinators and affiliation]. You have been invited to share your opinions about Internet Security.

Procedures

If you agree to participate in this research study, the following will occur: You will be asked to provide some demographic data about yourself. No personally identifiable information will be gathered. You will be asked about your experience with computers. You will be asked to comment on your experience and feelings regarding Internet security. The interview will be audio recorded to ensure accuracy in reporting your statements. The entire study should take about one hour.

Risks/Discomforts and Benifits

If you experience any discomfort, you may stop the study at any time. There are no direct benefits to you for participating in this study.

Confidentiality

The audio recording of this study will be transcribed to computer and then destroyed. All research data will be kept on a password-protected computer in a keypad-locked room on the [storage location]. Only the researchers will have access to this data. A unique, random ID will be generated for each study participant, and this ID will be used in place of any personally identifying information. Data will largely be presented in aggregate, but when direct quotes are required, they will be provided alongside the associated ID and will not contain personally identifying information. We may share research data on the Internet, but will not include any personally identifying information with this data, only the unique, random ID.

Compensation and Participation

You will be compensated $25 for your participation. Participation in this study is entirely voluntary. You have the right to withdraw at any point during the study or to refuse participation entirely. If you withdraw before the end of the study, you will still receive the full $25 compensation.

Questions about the Research

If you have any questions about this study, feel free to contact any of the following: [contact info]

Questions about Your Rights as a Research Participants

If you have questions regarding your rights as a research participant contact IRB Administrator at [contact info].

Statement of Consent
I have read, understood, and received a copy of the above consent and desire of my own free will to participate in this study.

Name (Printed): ______________________
Signature: ______________________
Date: ______________________

A.2 Demographic Handout

What is your gender?
- ○ *Male*
- ○ *Female*
- ○ *I prefer not to answer*

What is your age?
- ○ *18 – 24 years old*
- ○ *25 – 34 years old*
- ○ *35 – 44 years old*
- ○ *45 – 54 years old*
- ○ *55 years or older*
- ○ *I prefer not to answer*

What is the highest degree or level of school you have completed?
- ○ *Some school, no high school diploma*
- ○ *High school graduate, diploma or the equivalent (for example: GED)*
- ○ *Some college or university credit, no degree*
- ○ *College or university degree*
- ○ *Post-secondary education*
- ○ *I prefer not to answer*

What is your marital status?
- ○ *Married*
- ○ *Single*
- ○ *Other*
- ○ *I prefer not to answer*

Do you have children?
- ○ *Yes*
- ○ *No*
- ○ *I prefer not to answer*

A.3 Interview Guide

Introduction
- "Hello, my name is [name]. I am a researcher from [institution]. Before we begin, we have this consent form for you to read and sign."
- "Here is a short demographic survey."
- "Our research group is trying to understand how security affects you when you use the Internet. Our goal is to design software that makes it easier for you to be secure while you are online."

  "The opinions and ideas you share during this survey will be used to direct the future work of our research group. As such, feel free to be frank and honest. If at any time you have a thought or a comment, feel free to share it, regardless of whether you think it directly impacts the current topic."

Understand the Computing Environment
- How familiar would you say you are with computers?
  - How long have you been using them?
  - Do you use them at work/school?
- How many computers do you own?
  - Use on a daily basis?
  - Mobile devices?
- What sorts of things do you do on the Internet?

Threats
- When you are using the Internet, what dangers are you most concerned about?
  - Do your concerns change when you are at home/work/school?
  - Are there any dangers that affect your immediate family, but not you?
- Have you ever personally suffered harm from the Internet?
  - What was the nature of the harm?
  - What did you do to resolve the problem?
- What do you do to protect yourself while using the Internet?
  - Why do you do this?
  - How effective do you think these methods are?
  - Which one is most important?
  - Have you ever been unable to do something for fear of potential harm?

Encryption
- Have you seen this lock icon in your browser before? (Figure 7)
  - What does it mean to you?
  - If your website tells you that your connection is secure, what does this mean to you?
  - How do you feel when a website says it is secure?
- Do you ever send sensitive information over the Internet? For example, email or Facebook?
  - What types of sensitive information do you send.
  - What are you concerned about when sending sensitive information over the Internet?
- When I say "encryption", what do you think?
  - If they have heard of it.
    * What does it mean to you?
    * How do you encrypt data?
    * What assurances does encryption give you?
    * How easy is it for an attacker to steal encrypted data?
  - If they haven't heard of it.
    * Encryption is a process by which data is protected so that only you and intended recipients can read that data."
- Would you be interested in encrypting data you store or send over the Internet?
  - What data would you use encryption for?

- What services would you want it available with?
- How often would you encrypt data?
- Who would you send encrypted data?
- Would you want all of your messages encrypted? Why?

Notifications

- There are many ways that your computer can notify you of potential security problems.
  - What types of notifications that you currently see do you like best?
  - What annoys you the most about current notifications you receive?
- What would your ideal notification be like?
  - Could you please sketch a picture of your ideal notification.
  - How certain should the computer be before notifying you of a problem?
  - How often should you get a notification?

TLS Warning

- Here is a picture of a warning that browsers sometimes show (see Figure 8). Have you seen a similar warning before?
  - What do you do when you see this warning?
  - Under what circumstances do you ignore the warning and click through?
  - Under what circumstances do you stop trying to go to the website?
  - How often do you need to get to the underlying website, regardless of the warning.
  - Have you ever wished these warnings would just go away?

Closing

- “That is all we have time for. Thank you for your participation.”

A.4 Figures

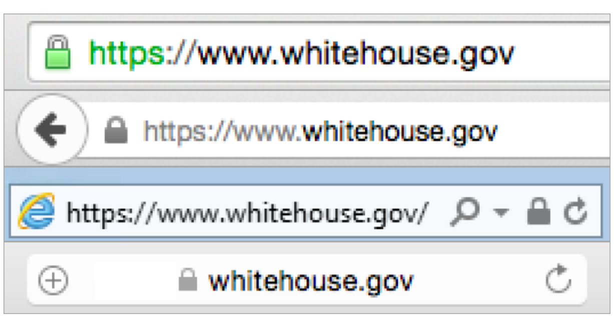


Figure 7: Example of lock icon from Chrome, Firefox, Internet Explorer, and Safari that were shown to participants.

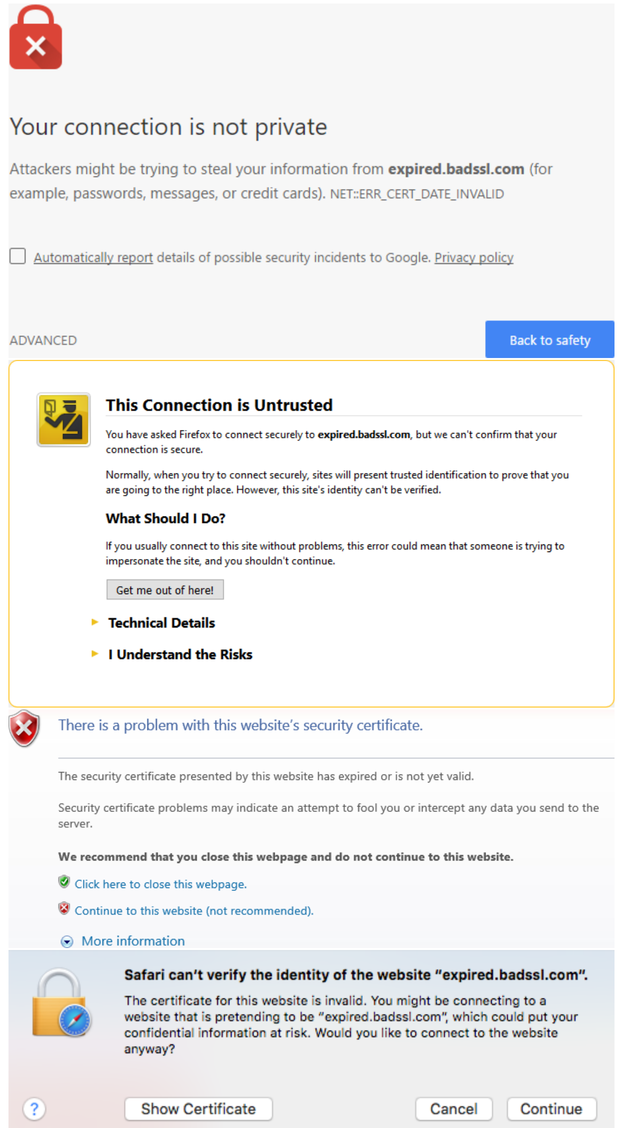
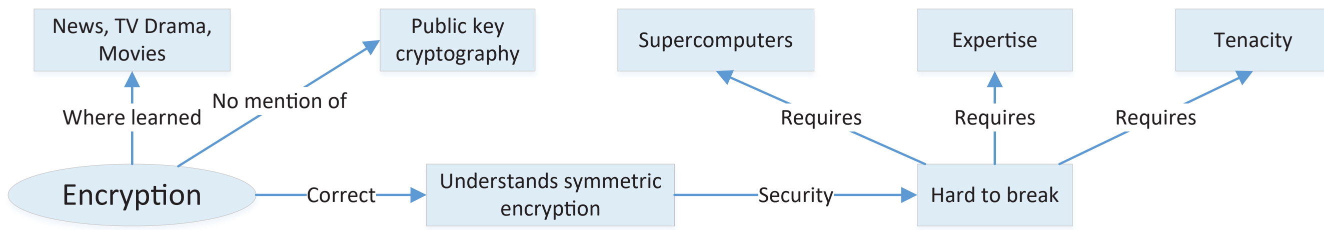


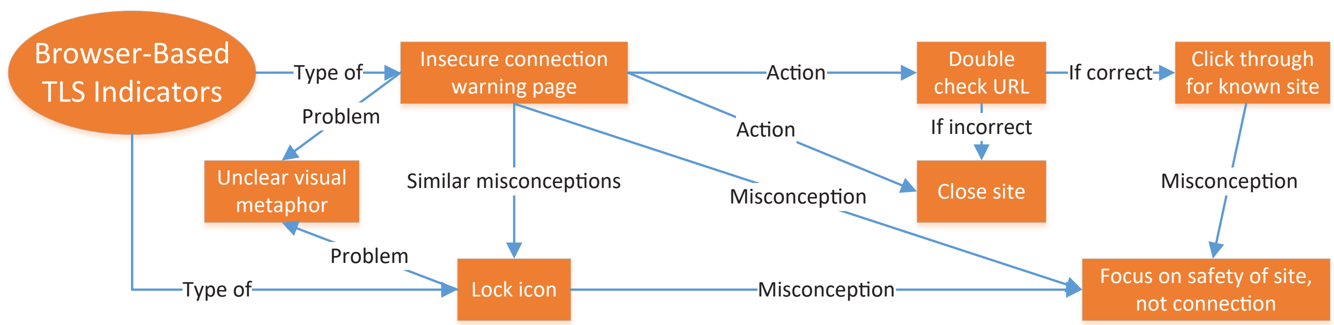
Figure 8: Example of TLS warnings from Chrome, Firefox, Internet Explorer, and Safari that were shown to participants.

B. ADDITIONAL CATEGORY GRAPHS



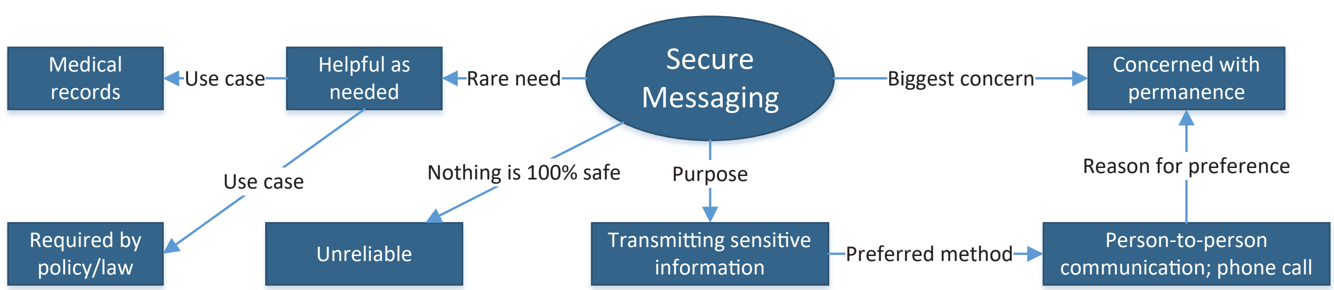
Participants learned about encryption from the news, TV dramas and movies. Most participants had a basic understanding of symmetric encryption, but almost no participants had knowledge of public key cryptography. In line with their belief that nothing is perfectly secure, participants noted that tenacious hackers could break encryption. This view accurately reflects the real world, as hackers consistently break systems that are "protected" by encryption.

Figure 9: Encryption Category Graph



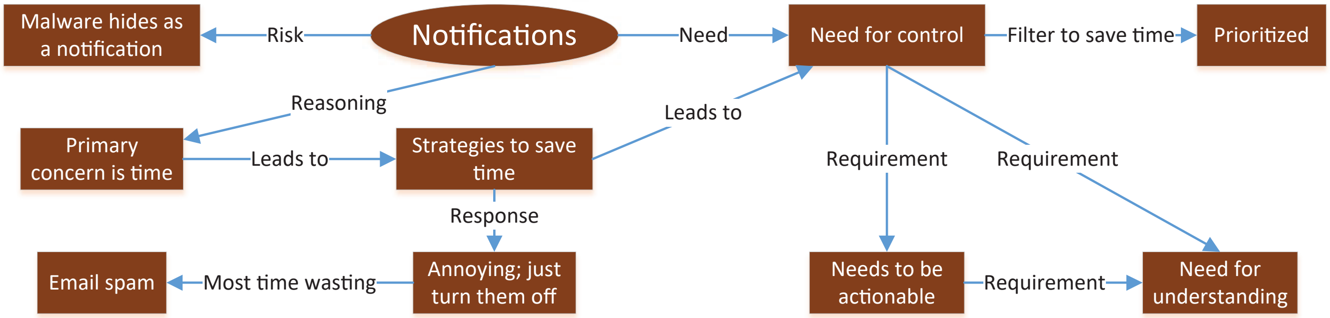
Many participants believed that the TLS lock icon and warning pages were related to site safety, and not the security of the connection. This misconception led them to ignore TLS warnings for well-known sites (e.g., Amazon) that they considered to be safe.

Figure 10: Browser-Based TLS Indicators Category Graph



Participants were interested in the potential of securing their online connections but were unsure whether this is even possible. They noted that the permanence of data from online communication (e.g., email) allows it to be attacked either during transmission or afterward. For this reason, they preferred to transmit information in person or over a phone call, which they viewed as non-permanent.

Figure 11: Secure Messaging Category Graph



Participants were largely apathetic towards notifications and warnings from security software (e.g., anti-virus). If viewed, participants wanted notifications to explain the problem to them, indicate the actions they could take, and explain the impact of those actions.

Figure 12: Notifications Category Graph

How Effective is Anti-Phishing Training for Children?

Elmer Lastdrager[1], Inés Carvajal Gallardo[1], Pieter Hartel[1,2], Marianne Junger[1]
[1]University of Twente, The Netherlands
[2]Delft University of Technology, The Netherlands
elmer@lastdrager.com, i.r.carvajalgallardo@utwente.nl, pieter.hartel@utwente.nl, m.junger@utwente.nl

ABSTRACT

User training is a commonly used method for preventing victimization from phishing attacks. In this study, we focus on training children, since they are active online but often overlooked in interventions. We present an experiment in which children at Dutch primary schools received an anti-phishing training. The subjects were subsequently tested for their ability to distinguish phishing from non-phishing. A control group was used to control for external effects. Furthermore, the subjects received a re-test after several weeks to measure how well the children retained the training. The training improved the children's overall score by 14%. The improvement was mostly caused by an increased score on the questions where they had to detect phishing. The score on recognizing legitimate emails was not affected by the training. We found that the improved phishing score returned to pre-training levels after four weeks. Conversely, the score of recognition of legitimate emails increased over time. After four weeks, trained pupils scored significantly better in recognizing legitimate emails than their untrained counterparts. Age had a positive effect on the score (i.e., older children scored higher than younger ones); but sex had no significant influence. In conclusion, educating children to improve their ability to detect phishing works in the short term only. However, children go to school regularly, making it easier to educate them than adults. An increased focus on the cybersecurity of children is essential to improve overall cybersecurity in the future.

1. INTRODUCTION

Fraudsters use phishing to convince victims to give out personal information. Commonly, the fraudsters want credentials that are used to access online services, such as online banking. Even though the impersonated brands that are misused in phishing are predominately financial institutions and payment providers, there has been a recent shift towards retailers and service-oriented companies [3, 4]. Several countermeasures are currently in use to prevent phishing victimization: blocking phishing messages and websites, improving interfaces, and training users [17].

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

Many training programs have focused on adults (e.g., [27, 5, 1, 18]). An often overlooked group of potential victims is children, with data about children only sparsely available (e.g., in [23]). The current generation of children, sometimes referred to as the *digital generation* or *digital natives*, grew up with the internet. The phrase "digital natives" is being criticized [6], since being a child in this generation does by itself not result in being more digitally capable. Instead, there are lots of opportunities for children, as well as adults, to use technology. Indeed, by the age of nine, many European children have access to the internet [15]. Many of the internet services that adults use, such as social media, email, or online gaming, are used by children as well [7]. A quarter of European children aged 9-10 and 73% of 13 to 14-year-olds have at least one profile on a social media website [15]. In the USA, 68% of teenagers aged 13-14 use social media [24]. Children, and in particular teenagers, are very well represented on the internet, with 92% of American children (13-17 years) [24] and 60% of European children (9-16 years) going online daily [15].

One might wonder why children are at risk. To illustrate why children could be targeted, consider the marketing domain. Marketers know that children have influence over what their parents buy and consequently target children in commercials [10]. In addition to marketing on TV, digital marketing offers even more chances to target children specifically [10, 28]. Phishing is commonly thought to be equivalent to theft of credentials of financial institutions. Since children often don't participate in online banking, what makes them attractive to a phisher? The online footprint of children on social media, websites, and email can be a target by itself. Obtaining access to email or social media accounts is valuable in order to access to a victim's network of friends and family. A phishing message that is sent by a friend is more likely to be opened than one from a stranger [18]. Subsequently, both children and adults within the victim's network can be approached with personalized phishing messages. Alternatively, influencing a child to provide the personal information of his or her parents provides helpful information for a follow-up call or email, even with simple pieces of information such as a phone number or home address. Training is needed to reduce the risk of initial victimization. Just like adults, children need to develop the ability to identify fraudulent communication, such as phishing emails.

Anti-phishing training can be administered in various ways. Advice can be given on an individual level, such as parents teaching their child how to ride a bike. Alternatively, one may educate a group at the same time; for example, schools teach skills like arithmetic to entire classes. When possible, educating a group of children can be more efficient. Since most children attend school, they are used to getting information in a class setting. Furthermore, when parents are insufficiently experienced to educate their children in the area of cybersecurity, this topic should be taught at school.

Education tackles only a part of the problem. An important issue is knowledge retention. One of the difficulties with user training is the extent to which the audience remembers the lessons over the long term. Retention indicates the effectiveness of training. Additionally, it is important to know how often to repeat training. This is true for traditional training, as well as alternative methods of creating user awareness, such as training by playing games [22, 34]. Studies performed on adults found no significant decay in performance from one week up to one month after the intervention [23, 21, 1, 27, 20]. This suggests that improvement of awareness after training is retained in the relatively short term. The question arises whether the same applies to children, as well as, more importantly, whether the improvement in awareness is stable over a longer period.

Children are very active online and can be the target of phishing e-mails. Accordingly, like adults, they should be trained to reduce the risk of victimization. This raises three questions to be answered. Firstly, what are children's abilities to detect phishing emails and websites? Secondly, what effect does cybersecurity training have on the children's ability to detect phishing? Thirdly, after receiving an awareness training, how well do children retain this knowledge? To answer these questions, we conducted empirical research.

Our contributions are: (1) to our knowledge, we are the first study to focus on the effect of anti-phishing training on children; (2) the training was based on storytelling and resulted in an improved detection of phishing in the short term and an improved detection of legitimate messages after 2–4 weeks; (3) we show that subjects with more online exposure, as well as older children, score better on a phishing identification test.

2. METHODOLOGY

An experiment was conducted at six schools in the Netherlands, using a cybersecurity training program that was designed for children aged 9–12. We tested their ability to recognize phishing and measured the effect of an intervention.

2.1 Design & Concepts

The experiment used a 2x2 between-group design. The training intervention was given on a group level (i.e., in a classroom), and we wanted to preserve the anonymity of the pupils. Therefore, no identifying information was recorded on the tests. Consequently, we did not record demographic data other than age and sex. The independent variables were the experimental condition (intervention or control) and the retest duration (measured in number of weeks). The outcome variable is the score on the test, ranging from 0 (no correct answers) to 10 (all answers correct). Five other variables were recorded to identify differences between groups and measure for certain individual differences: sex (male/female); age; possession of email address (yes/no); possession of a Facebook account (yes/no); and whether the subject had received a phishing email before (yes/no/unknown).

We will briefly discuss why these variables were included. Firstly, the subject's sex (male/female) was recorded because several phishing studies found that men are less prone to phishing victimization than women [5, 33, 18, 23], though other studies found no relationship [2, 13, 25]. Age was recorded with the expectation that older subjects would outperform younger ones [23, 33, 2]. Finally, the Routine Activity Approach states that for a crime to occur, a target and an offender must converge in the absence of a capable guardian [12]. Consequently, we expected children who are more active online to be more exposed to phishing. Therefore, subjects were asked whether they possess their own email address and Facebook account, and whether they have received a phishing email in the past.

In this paper, we use the terms "children" or "pupils" interchangeably to refer to the subjects of the study. "Teacher" refers to the school teacher of the pupils. The trainer is a researcher performing the study (by giving the presentation).

To establish the effectiveness of the cybersecurity training, we formed two types of groups: *intervention* and *control*. The intervention group was made up of school classes that received the cybersecurity training, followed by a capability test. To evaluate the effectiveness of the training, we compared the intervention group with a control group that received training after the study was finished (see Section 2.2).

2.1.1 Training and Procedure

A cybersecurity training program was developed for this experiment, consisting of an interactive presentation and a test. During the 40-minute presentation, the trainer would introduce and discuss cybersecurity with a class of pupils. The trainers were researchers and master's students specializing in cybersecurity. Asking children for their attention during a presentation can be challenging. Storytelling is an efficient method for non-experts to share in an expert's knowledge [31]. Therefore, the trainer used short stories and examples focussed on children to attract their attention.

The presentation provided the children with the necessary means of recognizing cyber misbehavior and advice on what to do. Topics included cyberbullying, hacking, phishing and identity theft. For phishing, we first explained what phishing is. Then, we showed an educational TV commercial that had been designed by the Dutch banking association [37]. Following the commercial, we asked the children in a group discussion what clues one should look for. Afterwards, we introduced four clues for identifying phishing emails: (1) how to find a URL from a hyperlink and how to assess where a URL leads to; (2) grammar, spelling, and the general type of language used; (3) presence of a sense of urgency or use of threats; and (4) the sender address. Furthermore, we showed two clues for websites: (1) the URL and (2) the need for an HTTPS connection when entering any data. During the training, the children were given ample opportunity to tell about their experiences, which helps the attendees remember the message. This led the children to share their own advice on how to prevent victimization, along with the advice that was included in the training. The trainer informed the children about the effectiveness of their own advice. Where needed, alternative advice was provided.

During the experiment, researchers went to schools in pairs. There were several practical constraints in time and availability. For example, schools had to book time to receive us, so there was a strict requirement to finish in time. Within classes of the intervention group, the trainers gave a presentation to the pupils. After the presentation, the children were given a paper-based phishing awareness test. Classes in the control group were only given the phishing test. No further explanation was provided, other than that the trainers would be back at a later time. Some pupils asked questions about a particular part of the test. The trainers answered that the pupil should pick the answer that made the most sense to the pupil.

After several weeks, each class was visited again. All pupils were given another paper-based phishing test. Finally, each child was given a one-page debriefing letter that explained and summarized the study. Additionally, all subjects were encouraged to discuss the test with their parents and contact one of the researchers with any questions.

2.1.2 Testing

Establishing the ability of children to detect phishing was measured using a paper-based phishing test. The participating schools did not have a computer available for each pupil. To allow school participation with the least effort, we chose a paper-based test over a computer-based test. The method of testing phishing ability and the introduction to the test can influence the results. For example, Parsons et al. [29] have shown that primed study participants are significantly better at discriminating between phishing and non-phishing compared to uninformed participants. To reduce this bias, children were not told that the goal was to discriminate phishing from non-phishing. Rather, the test was introduced as a 'cybersecurity test.'

The phishing test consisted of 10 questions, with six emails and four websites to judge. Both legitimate and phishing emails and websites were included. One correct answer was worth a point, and number of correct answers was the student's score on the test. Answering everything wrong would give a score of 0; answering everything correctly gave a 10. For each email or website in the test, a decision had to be made whether or not to take action. Although it was not stated explicitly, the pupils made a phishing or not phishing decision. Participating pupils were asked to note what kind of action they would take. Subjects' scores can vary depending on the type and origin of emails they have to judge [29]. Therefore, diversity in the types of emails and websites is essential to obtain a valuable result. Each question contained a clue as to why it should or should not be trusted. Some clues were explicit, such as a wrong link in an email or an unusual sender address. Others were based on the content, such as expressing urgency and spelling errors. For content-based clues, we made sure to include several in an email or website. All clues were mentioned in the training. The questions, emails, and websites were tailored to children and included a variety of different companies, such as toy stores, TV programs, game websites, a bank, and social media. The questions were not based on real-life phishing emails, since we are unaware of phishing attacks that target children specifically. However, we used existing legitimate emails and websites and adapted them, just like a phishing offender would do.

The tests were aimed at measuring the ability to identify emails and websites as phishing or legitimate correctly. However, using the same phishing test for the initial measurement as well as the re-test could result in the subjects remembering the questions. To avoid this memory effect, three sets of questions were used to measure the ability of children to detect phishing emails and websites. Three versions of the test were made: A, B, and C. Tests A and B included a front page with questions about the online exposure of the subjects. Test C was used in the pilot phase of the experiment and contains reordered questions from Test A.

Each subject got an overall score, the outcome variable. However, human beings generally assume that a message is truthful, and have great difficulty recognizing lies [26]. This has been called the truth bias [19, 26, 9]. We need to consider two parts in the subjects' performance: detecting lies (phishing) and detecting truth (legitimate). To do so, we made two equal-sized sets of questions. One set contained phishing, the other contained legitimate communications. By separately grading both sets of questions, we could distinguish between the ability to detect lies versus the ability to detect the truth. The overall score of a subject was calculated as the sum of both sets. For example, if a subject scored 3.0 out of 5 for recognizing phishing, and 2.5 out of 5 for recognizing a legitimate communication, the overall score would be 5.5 out of 10.

2.1.3 Retention

To measure knowledge retention, each school class took two phishing tests to test their ability to recognize phishing over time. Classes in the intervention condition received the training, followed by a test. Immediately after groups in the intervention condition finished their tests, the correct answers were discussed in class. This allowed the children to ask questions once more and get feedback on their decisions, thereby increasing the learning effect. After either 2 weeks (14 days), 4 weeks (28 days), or 16 weeks (64 days) a second test was done. Classes in the control condition did one test initially, followed by a re-test after 2 or 4 weeks. For the control condition, the results of the tests were not discussed in class. Unfortunately, classes in the control group that were scheduled for a re-test after 16 weeks were unable to participate the second time. This makes it impossible to compare the intervention group with a control group at 16 weeks. Therefore, our analysis will focus on the retention between 0 and 4 weeks.

2.2 Ethics

As with any experiment with humans, ethics are important. First of all, the design of this study was approved by the institutional review board of the University of Twente. The study was designed such that the subjects were not hurt or distressed in any way. Furthermore, each participating school was asked for permission to conduct the training and test their pupils. Additionally, we asked each participating school to distribute informed consent letters to the parents of their pupils. Parents were asked to sign and return the informed consent, either to the school or by email to the researchers. The contact information of the researchers was included in the informed consent, in case parents had questions. Several parents contacted the researchers. Only when the parents of a pupil had signed the informed consent and returned this to the school could a child participate as a subject.

After finishing the experiment, each subject was given a debriefing letter. The letter was written for the child and encouraged him or her to discuss the training with his or her parents. Furthermore, the contact details of the researchers were included in the debriefing, in case anyone had questions. After finishing the experiment, nobody contacted the researchers with questions.

From the point of view of the experiment, it was important to separate intervention and control groups. We considered it unethical to deprive subjects in the control group of a cybersecurity training. Therefore, after finishing their second phishing test and concluding their participation as subjects, pupils in the control group received the training too.

2.3 Setting

The experiment was held at six schools in the Netherlands, of which five primary schools and one secondary school. Each participating school gave permission for two sessions for at least one class. Every class received two tests (of 20-30 minutes each), and one intervention (about 40 minutes). Classes were randomly assigned to either an intervention group or a control group, and were additionally assigned a retention period by the researchers. All tests were taken individually by the subjects. The researchers were present to answer questions, but would never give away the correct answer. The subjects were told to answer what they would do if they had received the email or visited the website.

2.4 Subjects

The subjects were 353 pupils from six participating schools. All subjects were aged between 8 and 13 (M=10.66; SD=1.05), and over half (54%) were female. Children could join the training only if their parents had given their written consent before the start of the program (refer to Section 2.2 for more information). Children who did not have permission from their parents were temporarily sent to another classroom. If changing rooms was not possible, non-participating children were moved to another part of the same classroom to work on another task. Each child was assigned to an intervention or control group, based on the class they were in. This resulted in 181 children in the intervention group who received training, compared with the control group consisting of 172 children. The re-test was taken by 177 children. We included the week 0 data for several classes that were unable to participate for the re-test. Specifically, the missing classes consisted of all control group classes for the 16-week re-test. This resulted in the exclusion of the 16-week intervention group's re-test, since we could not compare them with their control group counterparts. Therefore, the number of subjects in week 0 is significantly higher compared to those for the re-tests in weeks 2 and 4. The exact number of subjects at each stage in the experiment is listed in Table 1.

2.5 Analysis

The three research questions guided the analysis. Descriptives of the control groups provided an answer to the first research question (i.e., what are the children's abilities to detect phishing emails and websites?). Furthermore, we tested whether the subject's characteristics influenced the score. An independent group t-test was used to measure the effect of the subject's sex and possession of an email account. The second research question was: what effect does cybersecurity training have on children's ability to detect phishing? To measure this effect, we compared the intervention group and the control group at 0 weeks. This was done using an independent group t-test, showing the difference between trained children (the intervention group) and untrained children (the control group). The third research question quantified the retention of the training. To answer this question, several linear regression models were developed. Firstly, a multi-level model was tested, measuring whether the school attended by the subject accounted for the results of the pupils. Even though the multi-level model was significant, the intraclass correlation was low (i.e., below 0.025). Therefore, linear regression was used instead. We developed several such models.

Table 1: Number of subjects in each stage of the experiment.

Group	Week 0	Week 2	Week 4
Intervention	181	49	38
Control	172	32	58

Model I uses the type of experiment (i.e., intervention or control), the number of weeks, and the interaction of these two as the predictors. ExperimentType shows the effect of the training on the score. The number of weeks indicates retention over time. Additionally, it is interesting to learn whether the effect of the training increases or decreases over time. For example, teaching someone a skill such as biking results in a higher level of skill over time if the person practices on his or her own. Therefore, the interaction between having participated in the intervention and the number of weeks (ExperimentType $\times$ Weeks) was taken into account as well. With this interaction, we could analyze whether the intervention resulted in better results as time progressed. A second model including social variables was constructed as Model II. Age and sex were added to the variables from Model I. Age was included since related literature suggested that older subjects score better than younger ones. The literature is inconclusive when it comes to sex and phishing victimization. Therefore, we added sex as a variable. Finally, Model III combines Models I and II and adds the test version and school, to show their potential influence on the overall score of the subjects. The school and test version variables were moderately correlated (r=0.68), as a consequence of Test C being used only in the pilot of the study. This results in collinearity in the model. Therefore, we omitted Test C from the model. These three models were used to predict the subject's overall scores on the tests.

Using the overall score as a measure of the ability to recognize phishing from legitimate is by itself insufficient. As discussed before, one needs to distinguish the differences in the scores of recognizing phishing and recognizing legitimate communications. To accommodate this, additional models were developed to distinguish lie detection and truth detection in the analysis. This lead to the introduction of six models. Phish-I through Phish-III were based on the previously described models I-III, but used the phishing (lies) score instead of the overall one. Additionally, Legit-I to Legit-III were developed to model the scores of the legitimate (truth) questions.

3. RESULTS

The first research question concerned the ability of children to detect phishing. This translates to the scores of the control group at the beginning of the experiment, at week 0. The average overall score of this control group is a 6.02 (Table 2) on a scale from 0 to 10. The overall score consisted of two parts: phishing (0–5 points) and legit (0–5 points). When considering only the questions that were related to phishing, the control group scores 3.74 on average, with a 95% confidence interval of [3.62, 3.88]. The mean score for labeling legitimate questions as such was lower: 2.26 (95% CI [2.09, 2.44]). In addition to the average scores of the control group, we also measured the effects of several subject characteristics on the overall score for all subjects. There was no significant effect of sex on the score, indicating a lack of evidence that boys performed differently from girls ($t(633) = -0.62$, $p=0.53$). There was a significant effect of age on the score, with older pupils scoring higher than younger ones ($F(1,633) = 6.28$, $p=0.01$, $R^2=0.010$, Adj. $R^2=0.009$). The effect of the school on the subject's score was significant ($F(5,636)=7.54$, $p<0.001$, $R^2 = 0.056$). One school scored significantly lower compared to the others ($B=-0.80$; $p=0.004$). Most of the subjects (80.3%) indicated having their own email address. Having one's own email address significantly influenced the score, with subjects having their own email address performing better than those without ($t(469)=3.68$, $p<0.001$). On the topic of social media, 26.6% of the subjects indicated having their own Facebook profile. Subjects with their own Facebook profile scored significantly higher than those without a Facebook profile ($z=2.330$, $p=0.02$, $r=0.10$). Thirdly, when asked whether they had ever received a phishing message, 8.9% answered 'yes', 37.4% answered 'no' and the remaining 53.7% responded that they did not know. Whether or not the subjects received a phishing email before was not significantly related to the subject's score ($F(2, 468) = 0.61$, $p=0.55$). A subject's online exposure did result in higher odds of having received a phishing message before ($F(2,215) = 6.25$, $p=0.002$, $R^2=0.040$), whereby having an email address was a significant indicator ($B=0.16$, $SE=0.05$, $p=0.04$).

To answer the second research question, the effect of the training was measured. Since three paper-based phishing tests were used in the experiment, we wanted the results to be comparable regardless of the version of the test. The mean overall results of pupils taking different tests were not significantly different from each other: A and B ($t(470)=1.89$; $p=0.059$); A and C ($t(307)=0.98$; $p=0.326$); B and C ($t(451)=-1.214$; $p=0.225$). Figure 1 shows the differences in scores in three box plots. The means and confidence intervals under all experimental conditions are listed in Table 2. The training itself resulted in an improvement in the scores of the participants in the intervention group that was statistically significant compared to the control group ($t(634)=-10.56$, $p<.001$). The effect size was $r=.39$, indicating a medium-sized effect [11]. In comparison, if we include only the first measurement (i.e., week 0), there is a significant difference between the untrained and the trained children as well ($t(351)=-5.19$; $p<0.001$). The training in week 0 had a small effect size of $r=.27$. These results show the effectiveness of adding a simple and short cybersecurity training to the curriculum of schools.

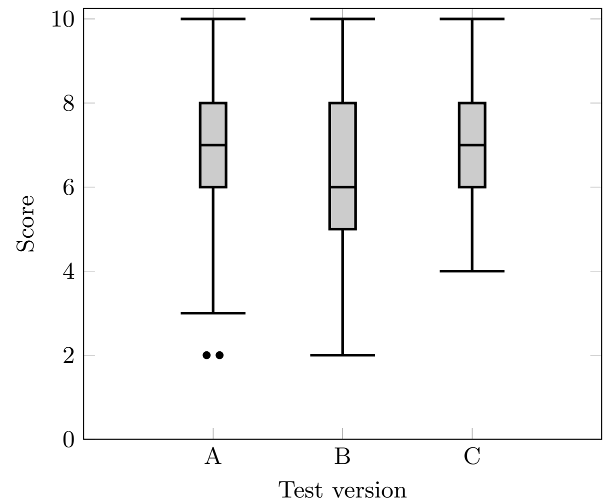


Figure 1: Box plot of three phishing tests of all observations (N=636).

To answer the third research question, retention over time was measured. Several linear regression models were constructed, the results of which are included in Table 3. Model I shows the influence of the cybersecurity training intervention on the score, as well as the effect over time, while controlling for the interaction effect. The resulting Model I is significant and explains 18.6% of the variance ($F(3,526) = 41.77$, $p<0.001$). Model II adds social predictors to Model I, resulting in a model that explains 19.8% of the variance ($F(5,523) = 27.63$, $p<0.001$). Finally, Model III includes the school as well as the version of the test, as well as the predictors from the other models. Model III is significant and explains 25.7% of the variance ($F(11,517) = 17.46$, $p<0.001$). In all three models, the effect of training significantly influenced the score of the subjects throughout the following weeks ($\beta=0.23$, $p<0.001$). Furthermore, the intervention group score significantly higher over time compared to the control group. Figure 2 plots Model III based on the number of weeks passed, split into intervention or control group, to show these effects visually.

To measure the differences in detecting lies from detecting truth, we developed additional models based on Models I, II and III. Instead of using the overall score as the outcome variable, we used the phishing score or the legitimate score, respectively. Since half of the questions were phishing, the scores range from 0 (all wrong answers) to 5 (all correct). Models Phish-I to Phish-III use the score of recognizing phishing. The model results can be found in Table 4. Model Phish-I includes the same predictors as the normal Model I, and is significant and explains 8.3% of the variance ($F(3,526)=15.36$, $p<0.001$). Model Phish-II is significant and explains 8.3% of the variance as well ($F(3,523)=9.26$, $p<0.001$). Model Phish-III is significant as well and explains 13.1% of the variance ($F(11,517)=9.60$, $p<0.001$). Compared to the models of the overall scores, different effects emerge. For example, subject age and weeks since inter-

Table 2: Mean score and 95% confidence interval per experimental setting.

Type	Week	Overall Score		Phishing Score		Legitimate Score	
		Mean	95% CI	Mean	95% CI	Mean	95% CI
Cont	0	6.02	5.79–6.26	3.61	3.45–3.77	2.41	2.20–2.62
Exp	0	6.87	6.65–7.09	4.26	4.15–4.38	2.61	2.41–2.80
Cont	2	5.72	5.21–6.23	4.09	3.74–4.45	1.62	1.17–2.08
Exp	2	7.95	7.58–8.34	4.33	4.12–4.53	3.63	3.28–3.99
Cont	4	6.14	5.75–6.53	3.96	3.70–4.23	2.17	1.79–2.55
Exp	4	8.13	7.67–8.60	4.00	3.73–4.27	4.13	3.81–4.46
Cont	all	6.01	5.82–6.20	3.74	3.62–3.88	2.26	2.09–2.44
Exp	all	7.35	7.19–7.51	4.23	4.15–4.32	3.11	2.97–3.26

Table 3: The linear regression models of the overall score.

Characteristic (reference)	Model I			Model II			Model III		
	B	SE B	β	B	SE B	β	B	SE B	β
ExperimentType (control)	0.92***	0.16	0.28	0.90***	0.16	0.27	1.00***	0.12	0.10
Weeks	0.01	0.06	0.01	0.03	0.06	0.03	0.11	0.12	0.10
Weeks × ExperimentType	0.34***	0.08	0.23	0.36***	0.08	0.24	0.30**	0.08	0.20
Age				0.18**	0.06	0.11	0.19**	0.07	0.12
Sex (female)				0.08	0.13	0.02	0.14	0.14	0.04
Test version (A)†									
– Test B							-0.17	0.39	-0.05
School (1)									
– 2							0.89**	0.33	0.16
– 3							0.44	0.31	0.08
– 4							-0.33	0.34	-0.05
– 5							0.30	0.43	0.07
– 6							-0.24	0.47	-0.07
Constant	5.99***	0.11		4.04***	0.69		3.80***	0.85	
R^2	0.186			0.198			0.257		
Model significance	0.000***			0.000***			0.000***		
N	530			529			529		

Note. Coefficients unstandardized (B) and standardized (β). SE=Standard Error. Significance (χ^2): * p<0.05; ** p<0.01; *** p<0.001. †Due to collinearity, the output of Test C was omitted.

vention in Phish-III are not significant, whereas they are in the overall Model III. The differences are more easily viewed when Model Phish-III is plotted in Figure 3a. At week 0, the intervention group's scores differ significantly from the control group, as shown by the confidence intervals. However, in week 4, there is no significant difference between the intervention group and the control group anymore. The control group scored similarly in week 4 compared to week 0. Subjects within the intervention group scored significantly lower in week 4 compared to week 0.

In additional to the three phishing-only models, three legit-only models were constructed. Similarly, three models, Legit-I to Legit-III were constructed based on the overall Models I to III, respectively. The results of these models can be found in Table 5. Model Legit-I was significant and explained 15.1% of the variance (F(3,526)=42.57, p<0.001). Model Legit-II was significant and explained 16.4% of the variance (F(5,523)=29.59, p<0.001). Model Legit-III was significant and explained 26.0% of the variance (F(11,517) =20.28, p<0.001). A graph showing Model Legit-III is included in Figure 3b, with scores ranging from 0 to 5 for all five questions testing legitimacy. There are no significant differences in score at week 0 between the intervention group and the control group for the legitimate scenarios (z=-1.17; p=0.24). In week 4, however, the scores of the intervention group and control group differ significantly (z=-5.85; p<0.001). During the experiment, the score of the control group did not change significantly (t(228) = 1.11; p=0.27). In the intervention group, a significant increase in score was observed between week 0 and week 4 (z=-6.05; p<0.001).

4. DISCUSSION

The concept of testing the ability to detect phishing in an educational setting is challenging [32]. Getting the attention of children aged 8–13 to focus on cybersecurity is no less of a challenge. Untrained children are mediocre at discriminating phishing emails and websites from legitimate ones, scoring 6.02 out of 10 in our experiment. However, subjects trained in a single 40-minute training session and interactive discussion scored 6.87 out of 10, an increase of 14% over their untrained peers. The overall score by itself is not sufficient as a measurement of accuracy, since humans are generally not very good at recognizing lies [26]. Therefore, we distinguished the correctness scores for phishing and legitimate questions.

Table 4: The linear regression models of the phishing-only score. The construction of the models is similar to Table 3.

	Model Phish-I			Model Phish-II			Model Phish-III		
Characteristic (reference)	B	SE B	β	B	SE B	β	B	SE B	β
ExperimentType (control)	0.65***	0.10	0.34	0.65***	0.10	0.34	0.70***	0.10	0.37
Weeks	0.10**	0.04	0.16	0.10**	0.04	0.16	0.02	0.07	0.04
Weeks × ExperimentType	-0.15**	0.05	-0.18	-0.15**	0.05	-0.17	-0.18**	0.05	-0.22
Age				0.01**	0.04	0.01	0.05	0.04	0.05
Sex (female)				-0.00	0.08	-0.00	0.09	0.08	0.05
Test version (A)†									
– Test B							-0.21	0.24	-0.10
School (1)									
– 2							0.56**	0.21	0.17
– 3							0.08	0.21	0.03
– 4							0.20	0.22	0.06
– 5							0.95**	0.27	0.41
– 6							0.77**	0.29	0.40
Constant	3.63***	0.08		3.50***	0.44		2.59***	0.52	
R^2	0.083			0.083			0.131		
Model significance	0.000***			0.000***			0.000***		
N	530			529			529		

Note. Coefficients unstandardized (B) and standardized (β). SE=Standard Error. Significance (χ^2): * p<0.05; ** p<0.01; *** p<0.001. †Due to collinearity, the output of Test C was omitted.

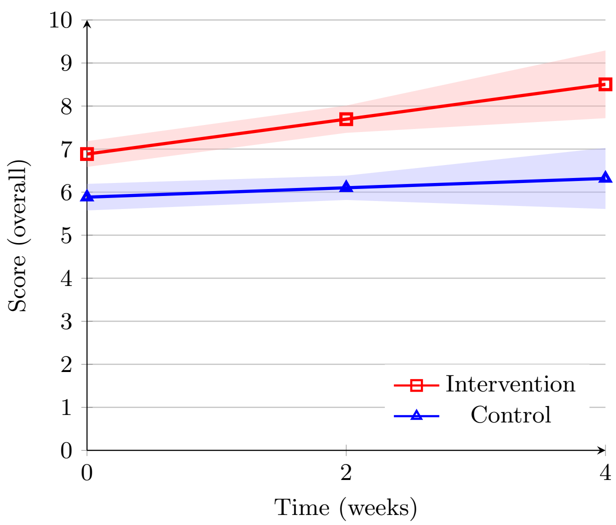


Figure 2: Overall predicted ability scores over time, in number of correct answers (0–10). Shades indicate 95% confidence interval. N=529.

We found that training improved the ability to recognize phishing directly following the training, but it did not significantly change the ability to identify legitimate emails correctly. This phenomenon has been discussed in the literature. Hauch et al. [16] have shown in a meta-analysis that training improves both overall accuracy and lie detection, but not truth detection accuracy. This was also the case in our experiment; the subjects did not score significantly better on truth accuracy of legitimate emails and websites on the test directly following the training, compared to the control group. This can be explained by the focus of our training on how to detect phishing. According to Hauch et al. [16], if the focus of training is on deception detection, the subject's post-training truth accuracy remains unaffected. An alternative explanation would be that the training made the subjects paranoid. However, if that were to be the case, the subjects would have to score lower on recognizing legitimate emails, which was not the case.

The overall scores of trained subjects improved significantly over time, indicating a good knowledge retention of the subjects. Within the control group, the overall scores remained stable. When considering only the phishing questions, subjects from the intervention group suffered from a small decay in their ability to recognize phishing. Specifically, after 4 weeks, the ability of the intervention group to recognizing phishing matched the level of the control group. Regardless of the decay over time, the scores on the phishing questions were relatively high, with averages of correct answers between 3.7 and 4.4 questions. Since 5 was the maximum, we believe that there is a ceiling effect: many subjects achieved the highest score, and could not improve their scores further. Our test consisted of 10 questions composed of two sub-tests, five legitimate and five phishing. This means that subjects could not receive higher scores than 5 on both sub-tests, which is the maximum on our measures. When many subjects have the maximum score, their scores cannot be distinguished. Figure 3b illustrates this clearly for the intervention group. Therefore, only less-performing subjects could increase their score after training. The subsequent score decay over time shows that the effect of the training, in terms of the ability to recognize phishing emails, fades within a month. To the best of our knowledge, no similar phishing tests have been undertaken with children, making comparisons with other phishing literature difficult. There are studies on phishing interventions with adult subjects, which found no significant decay of the trained subject's abilities after 7 to 28 days [23, 21, 1, 27, 20]. However, there are major methodological differences, since the above-mentioned studies use interactive, computer-based methods of training, such as playing games [23, 21, 27] or roleplay-

Table 5: The linear regression models of the legitimate-only score. The construction of the models is similar to Table 3.

	Model Legit-I			Model Legit-II			Model Legit-III		
Characteristic (reference)	B	SE B	β	B	SE B	β	B	SE B	β
ExperimentType (control)	0.27	0.14	0.09	0.25	0.14	0.09	0.30*	0.14	0.10
Weeks	-0.08	0.05	-0.09	-0.07	0.05	-0.07	0.08	0.11	0.09
Weeks × ExperimentType	0.49***	0.07	0.38	0.51***	0.07	0.39	0.48***	0.07	0.37
Age				0.17**	0.06	0.11	0.14*	0.06	0.10
Sex (female)				0.08	0.12	0.03	0.05	0.12	0.02
Test version (A)†									
– Test B							0.04	0.35	0.01
School (1)									
– 2							0.33	0.29	0.07
– 3							0.36	0.26	0.07
– 4							-0.54	0.29	-0.10
– 5							-0.65	0.40	-0.18
– 6							-1.02*	0.41	-0.35
Constant	2.36***	0.11		0.54	0.62		1.21	0.74	
R^2	0.151			0.164			0.260		
Model significance	0.000***			0.000***			0.000***		
N	530			529			529		

Note. Coefficients unstandardized (B) and standardized (β). SE=Standard Error. Significance (χ^2): * p<0.05; ** p<0.01; *** p<0.001. †Due to collinearity, the output of Test C was omitted.

ing [1]. However, within the field of social engineering, it has been reported that an intervention to increase awareness is subject to significant decay [8], showing social engineering awareness returning to pre-intervention levels after two weeks.

While the phishing score decreased over time, the score for legitimate questions followed a rather different pattern. The score over time increased significantly, contrary to our expectations. After two and after four weeks, subjects in the intervention group were able to correctly recognize legitimate scenarios significantly better than subjects from the control group. The cybersecurity training may have triggered the interest of the children, causing them to pay more attention to messages they receive, or to think about the lessons learned. Another possible explanation is that the subjects trained themselves based on emails they received in their daily lives. This may be compared to learning how to ride a bike, where an initial set of skills and knowledge is needed to start biking, and with more practicing, performance increases over time. In other words, training made the children look more closely at the emails they received, after which they were better at identifying legitimate emails.

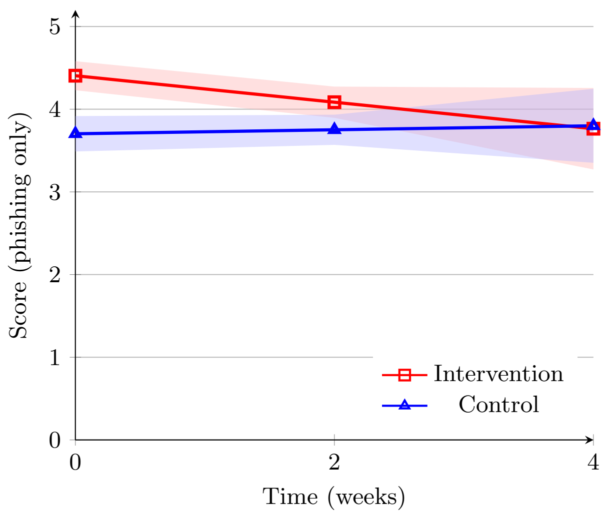
Further trainings, sometimes called boosters, could be used to increase these abilities and counter decay of the ability to recognize phishing [20, 30]. However, regular training is costly. In the context of children, it may be infeasible for schools to introduce boosters on a regular basis. This is especially the case when the retention of knowledge is short (i.e., a month). Training using different methods, such as letting the subjects play a game [23, 21], may be less affected by this disadvantage since the subjects can play the game regularly without supervision. Before introducing additional training, however, better measurements should be used to identify the problem better. One possible fix is an extensive test with more questions and more challenging questions, which could be used to avoid a possible ceiling effect. That way, subjects would be less likely to get the maximum score, and decay or increase effects should be more visible.

Another finding is that older children score better than younger ones. This is in line with similar studies about phishing interventions on adults. In several studies, young adults perform worse than older ones [33, 2]. In particular, a large-scale study [23, 33] found that teenagers between 13 and 17 perform worse than adults in phishing tests. A possible reason for this result is lower education and fewer years of internet experience [33]. Furthermore, subjects in this study who have their own email address or a Facebook profile scored significantly higher than other subjects. This suggests that, indeed, internet experience may be an influential factor. Another factor that could influence the subject's score is the training itself. Despite efforts to make all trainings similar, there are group dynamics involved, especially when relying on interaction (e.g., stories) with the subjects.

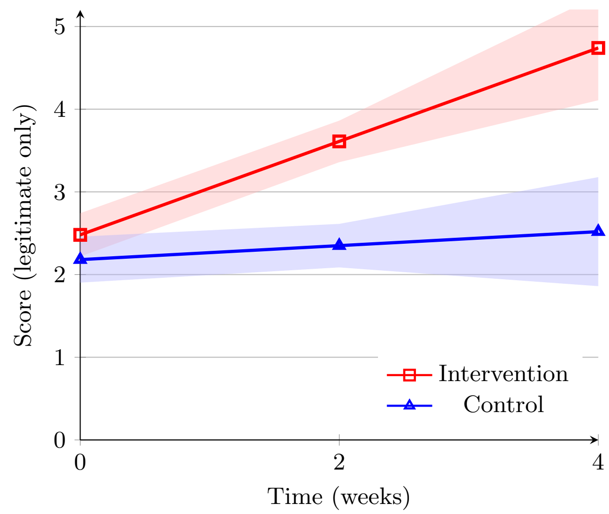
Other candidate relations did not significantly contribute to the final score of a child. In particular, the sex of the child had no significant influence on the score, when controlling for other variables. Specifically for children, sex differences are not necessarily to be expected at all. For example, boys only begin to take more risks than girls between the ages of 9 and 11 [35]. The lack of differences could be explained by the age groups of the children that participated. Additionally, even for adults and adolescents, the existence of a relation between sex and phishing knowledge is doubtful in existing literature [2, 13, 25]. The interaction between age and sex did not predict phishing knowledge of children either.

4.1 Limitations

There are several limitations to the results of this study. Even though the intervention condition was given per class, this did not prevent children in one class from talking to their peers in other classes. Since all parents were informed and asked for permission beforehand, they could have dis-



(a) Includes only the phishing questions.



(b) Includes only the legitimate questions.

Figure 3: Predicted ability score split by phishing and legitimate. Shades indicate 95% confidence interval. N=529.

cussed the topic of cybersecurity with their children. Unknown external factors may be responsible for the increase over time. For example, the participating children may have seen one of the phishing awareness commercials on TV. Personal experience of the researchers was that indeed one of of these three explanations was plausible. One of the colleagues at the University of Twente, who was not involved in the study, had a child in the intervention condition. The colleague mentioned that his children and the other parents were enthusiastic about the intervention and that he had talked about it at home. This example could explain the increase in ability over time that was observed. Moreover, this colleague had other children in the same school. Hence, the intervention could have influenced children in the control condition. However, we do not see indications of that effect in the data.

A possible critique on the study is that the children know that they are being tested. The results, therefore, do not necessarily reflect their ability when receiving an email in the wild. While this is true, we consider the tests an appropriate way to measure the subject's ability to recognize phishing. The subjects' scores are arguably different from how they would respond to a phishing email in their own inbox, since more factors are involved. Factors such as language (an eight-year-old Dutch child receiving an English email) and expectancy (not having a bank account) could increase their real world score. On the other hand, factors like attention (doing other things in parallel) and limited interfaces (not being able to check the link on a tablet computer) could affect resilience in the real world. Furthermore, the subjects received a second test a period of time after the first. This means that they know what to expect when they start the second test.

This study may suffer from an assignment bias. Even though the groups were assigned at random to one of the conditions, the number of schools that participated is limited. Furthermore, all schools are located in two cities in the east of the Netherlands. The results might be affected by factors unknown to the researchers. A nation-wide study on randomly selected schools could counter such biases regarding region and quality of teaching.

A presentation (or lecture) is one way to deliver a message to pupils. Other ways of teaching may be more efficient, such as using games [14]. We chose a traditional presentation-based intervention because it is relatively simple to apply to current primary schools. The pupils do not need to have access to a computer, and a presentation and paper-based test fit in well with the rest of the daily program and activities. Alternatively, game-based anti-phishing solutions [22, 34] may yield better results and could have different retention properties.

Using a paper-based test with images raises questions regarding the representativeness of the resulting score compared to real-world phishing. Whereas using static images or screenshots is not optimal, they have been used before in phishing experiments [36, 29, 33]. We believe there is little difference between seeing an image on a screen or seeing one printed on paper. Furthermore, not all subjects may be equally computer literate, and using static images on paper results in a level playing field.

Finally, all students filled in the tests anonymously. Therefore, no repeat measurements were available at an individual level. The analyses could therefore not be performed on repeated-measures samples. Rather, we treated the test results as independent samples. As a consequence, the reported results are conservative and an underestimation, as they miss the power of a repeated-measures test.

5. CONCLUSIONS

Children need to understand digital risks to reduce the risk of victimization on the internet. Understanding digital risks is important for children as well as adults. However, the majority of children are self-taught when it comes to the internet [7], making it unlikely they will systematically learn how to act safely. To learn about the abilities of children in detecting phishing emails and websites, researchers had children aged 8–13 take in a phishing recognition test. Half of the children received training before the test, and the other half did not. Both trained and untrained children were tested for the ability to distinguish phishing emails and websites from legitimate ones. Several schools participated in the study. A first indicator of the practical need for such training arose while performing the experiment. During the training, as more pupils started sharing their stories, they became very enthusiastic and asked lots of questions. In most classes, at least one child knew a phishing victim. These victims were mostly relatives or neighbors. The most common situation in the stories that were told was a victim losing money due to filling in banking credentials on a phishing website. Hearing stories from their peers impacted the children and provided them with a warning message stronger than the presenters could ever give.

Until novel anti-phishing techniques are developed and deployed on a large scale, user training seems to be important. For adults as well as children, that means creating an improved knowledge of the subject for as many individuals as possible. In many countries, all children aged 9 or older attend some form of education. Potentially, this makes it feasible to embed a cybersecurity training in their curriculum, effectively training the entire population of children.

In our experience, both schools and parents are very willing to embed lessons about cybersecurity in the curriculum. Our request to give a training was well received. In particular, incidents with phishing, cyberbullying, and other cyberthreats are often in the news. Teachers and parents reported being worried about those issues. At the same time, teachers at schools where we gave a training, found the course highly informative for themselves as well. Techniques for establishing the validity of an email were unknown to them. Several teachers mentioned that hovering over a hyperlink or checking the sender address were valuable approaches for them. Training teachers should, therefore, be the first step in cybersecurity education. Where needed, universities and practitioners (e.g., IT security firms) could provide help. There are existing initiatives, such as the $(ISC)^2$ Safe and Secure Online[1] where security professionals visit schools. Such initiatives should be extended to more countries and expanded in size, and new ones should be developed.

Training children increased their short-term ability to distinguish phishing from legitimate correctly. Specifically, their ability to recognize phishing increases significantly after an in-class training. However, this increased ability is subject to decay. After four weeks, the ability to recognize phishing for trained children diminished to the level of their non-trained counterparts. This suggests that the training created knowledge, but that this knowledge only lasted through the short term. On the positive side, trained children did continue to perform better in recognizing legitimate emails as such. This increases the odds of legitimate communications reaching the end user. Increasing the ability to recognize phishing requires good awareness.

All in all, we believe that researchers and practitioners in the field of cybersecurity should not only focus on adults, but that material for children should be developed in parallel. Phishing, specifically, is too often seen as an adult-only crime. The children of today are the victims of the future.

[1]See also `https://iamcybersafe.org/`

6. ACKNOWLEDGMENTS

We would like to thank Brinda Badarinath Hampiholi, Joey de Vries, Lorena Montoya, and Jan-Willem Bullée for their valuable advice and feedback. We would also like to thank the reviewers for their constructive feedback and shepherd Elizabeth Stobert for her helpful comments.

7. REFERENCES

[1] A. Alnajim and M. Munro. An Evaluation of Users' Anti-Phishing Knowledge Retention. In *International Conference on Information Management and Engineering*, ICIME '09, pages 210–214. IEEE, 2009.

[2] I. M. Alseadoon. *The Impact of Users' Characteristics on Their Ability to Detect Phishing Emails*. Phd thesis, Queensland University of Technology, 2014.

[3] Anti-Phishing Working Group. Phishing activity trends report, 3rd quartile 2014, 2014. `http://docs.apwg.org/reports/apwg_trends_report_q3_2014.pdf`.

[4] Anti-Phishing Working Group. Phishing activity trends report, 4th quartile 2014, 2015. `http://docs.apwg.org/reports/apwg_trends_report_q4_2014.pdf`.

[5] M. Blythe, H. Petrie, and J. A. Clark. F for fake: Four Studies on How We Fall for Phish. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 3469–2478, New York, NY, USA, 2011. ACM.

[6] d. boyd. *It's Complicated: The Social Lives of Networked Teens*. Yale University Press, 2014.

[7] C. Brady. Security Awareness for Children. Technical Report RHUL–MA–2010–05, Royal Holloway, London, 2010.

[8] J.-W. Bullee, L. Montoya, M. Junger, and P. Hartel. Telephone-based social engineering attacks: An experiment testing the success and time decay of an intervention. In *Singapore Cyber Security R&D Conference, SG-CRC 2015*, pages 1–6, Singapore, 2016. IOS Press.

[9] J. K. Burgoon and T. R. Levine. Advances in deception detection. In S. W. Smith and S. R. Wilson, editors, *New directions in interpersonal communication research*, pages 201–220. Sage, 2010.

[10] S. L. Calvert. Children as consumers: Advertising and marketing. *Future of Children*, 18(1):205–234, 2008.

[11] J. Cohen. A power primer. *Psychological Bulletin*, 112(1):155–159, 1992.

[12] L. E. Cohen and M. Felson. Social change and crime rate trends: A routine activity approach. *American Sociological Review*, 44(4):588–608, 1979.

[13] R. Dhamija, J. D. Tygar, and M. Hearst. Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, CHI '06, page 581, New York, New York, USA, 2006. ACM Press.

[14] A. Domínguez, J. Saenz-de Navarrete, L. De-Marcos, L. Fernández-Sanz, C. Pagés, and J.-J. Martínez-Herráiz. Gamifying learning experiences: Practical implications and outcomes. *Computers & Education*, 63:380–392, 2013.

[15] L. Haddon and S. Livingstone. EU Kids Online: national perspectives. Technical report, The London School of Economics and Political Science, 2012.

[16] V. Hauch, S. L. Sporer, S. W. Michael, and C. a. Meissner. Does Training Improve the Detection of Deception? A Meta-Analysis. *Communication Research*, pages 1–61, 2014.

[17] J. Hong. The state of phishing attacks. *Communications of the ACM*, 55(1):74–81, 2012.

[18] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer. Social phishing. *Communications of the ACM*, 50(10):94–100, 2007.

[19] D. Kahneman. *Thinking, Fast and Slow*. Penguin Books UK, 2012.

[20] P. Kumaraguru, J. Cranshaw, A. Acquisti, L. Cranor, J. Hong, M. A. Blair, and T. Pham. School of phish: A Real-World Evaluation of Anti-Phishing Training. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, New York, New York, USA, 2009. ACM Press.

[21] P. Kumaraguru, Y. Rhee, S. Sheng, S. Hasan, A. Acquisti, L. F. Cranor, and J. Hong. Getting users to pay attention to anti-phishing education: Evaluation of retention and transfer. In *Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit*, eCrime '07, pages 70–81, New York, NY, USA, 2007. ACM.

[22] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong. Lessons from a real world evaluation of anti-phishing training. In *2008 eCrime Researchers Summit*, pages 1–12. IEEE, 2008.

[23] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong. Teaching Johnny not to fall for phish. *ACM Transactions on Internet Technology*, 10(2):1–31, 2010.

[24] A. Lenhart. Teens, Social Media and Technology Overview 2015. Technical report, Pew Research Center, 2015.

[25] E. R. Leukfeldt. Phishing for Suitable Targets in The Netherlands: Routine Activity Theory and Phishing Victimization. *Cyberpsychology, Behavior, and Social Networking*, 17(8):551–555, 2014.

[26] T. R. Levine, H. S. Park, and S. a. McCornack. Accuracy in detecting truths and lies: Documenting the "veracity effect". *Communication Monographs*, 66(2):125–144, 1999.

[27] C. B. Mayhorn and P. G. Nyeste. Training users to counteract phishing. In *Proceedings of the Human Factors and Ergonomics Society*, volume 41, pages 1956–1960, 2012.

[28] K. C. Montgomery, J. Chester, S. A. Grier, and L. Dorfman. The New Threat of Digital Marketing. *Pediatric Clinics of North America*, 59(3):659–675, 2012.

[29] K. Parsons, A. McCormac, M. Pattinson, M. Butavicius, and C. Jerram. The design of phishing studies: Challenges for researchers. *Computers & Security*, 52:194–206, 2015.

[30] S. Purkait. Phishing counter measures and their effectiveness – literature review. *Information Management & Computer Security*, 20(5):382–420, 2012.

[31] E. Rader, R. Wash, and B. Brooks. Stories As Informal Lessons About Security. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, SOUPS '12, pages 6:1–6:17, New York, NY, USA, 2012. ACM.

[32] S. A. Robila and J. W. Ragucci. Don't be a phish: Steps in User Education. In *Proceedings of the 11th annual SIGCSE conference on Innovation and technology in computer science education*, ITICSE '06, pages 237–241, New York, NY, USA, 2006. ACM.

[33] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. S. Downs. Who falls for phish? A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 373–382, New York, New York, USA, 2010. ACM Press.

[34] S. Sheng, B. Magnien, P. Kumaraguru, A. Acquisti, L. F. Cranor, J. Hong, and E. Nunge. Anti-Phishing Phil. In *Proceedings of the 3rd symposium on Usable privacy and security*, SOUPS '07, pages 88–99, New York, New York, USA, 2007. ACM Press.

[35] P. Slovic. Risk-Taking in Children: Age and Sex Differences. *Child Development*, 37(1), 1966.

[36] A. Tsow and M. Jakobsson. Deceit and deception: A large user study of phishing. Technical Report TR649, Indiana University, 2007.

[37] Veilig Bankieren (Dutch Banking Association). Nepmail, daar trapt u niet in, 2011. TV commercial (Dutch): `http://youtu.be/VcbHoOEOtkA`.

"I feel stupid I can't delete...": A Study of Users' Cloud Deletion Practices and Coping Strategies

Kopo M. Ramokapane
Security Lancaster Institute
Lancaster University, UK
k.ramokapane@lancaster.ac.uk

Awais Rashid
Security Lancaster Institute
Lancaster University, UK
a.rashid@lancaster.ac.uk

Jose M. Such
Department of Informatics
King's College London, UK
jose.such@kcl.ac.uk

ABSTRACT

Deletion of data from cloud storage and services is an important aspect of privacy and security. But how easy or simple a task is it for users to complete? Cloud users' deletion practices, challenges and coping strategies have not been well studied to date. We undertook an exploratory study to better understand this issue. Through in-depth semi-structured interviews and use of deletion scenarios with 26 subjects, we explored several key questions: why and when cloud users would like to delete, why cloud users cannot delete, what causes such failures, what users do to work around these problems, and finally what do users want in terms of usable deletion in the cloud. We found that users' failure to delete arises from lack of information about deletion, incomplete mental models of the cloud and deletion within the cloud, and poorly designed user interfaces for deletion functions. Our results also show that users develop different coping strategies such as deleting from certain devices only, seeking help and changing service providers, to overcome such challenges. However, these strategies may not always produce desired results. We also discuss potential ways to improve the usability of deletion in the cloud.

1. INTRODUCTION

Since the advent of cloud computing incomplete deletion of data has been a concern for most organizations and users. Researchers have looked into provision of assured deletion in the cloud [6, 26] and encryption-based solutions to securely dispose of data after use [25, 29]. However, such approaches start from the assumption that users *know* data management in the cloud, have clear mental models of how deletion may operate in the cloud and can accomplish the task of deletion through either the features offered by cloud providers or using more sophisticated assured deletion mechanisms such as encryption-based solutions.

In this paper, we focus on the user's perspective and investigate usability of data deletion from cloud storage and services. We explore several key questions fundamental to usable privacy and security: what are the motivating factors that underpin cloud users' need to delete? Do they find current cloud deletion mechanisms usable and, if not, what are the factors underpinning users' failure to delete? What are the coping strategies that users deploy to work around these problems and what do users want in terms of usable deletion in the cloud?

Recent high profile incidents have highlighted the security and privacy concerns of users with regards to data management and retention in the cloud. For instance, Dropbox users were alarmed when their files and folders deleted as far as 5 years ago mysteriously re-appeared in their accounts [11]. Similar concerns have been raised by iCloud users upon learning that Apple had been retaining their browsing history for more than a year – several months after it was supposed to have been deleted [13].

There has been a substantial body of knowledge and experimental evidence on users' security and privacy behaviors, for instance [1, 30]. The usability issues of encryption mechanisms have also been well-documented, e.g., [10, 33]. Recent research on deletion in the cloud has focused on risks associated with incomplete deletion or retained data in the cloud [26] as well as encryption-based deletion solutions [25, 29]. However, the usability of data deletion in the cloud has not been explored and users' understanding and challenges of deleting from the cloud are still to be studied. For instance, at present, cloud users are allowed to access and delete from the cloud through mobile apps, web interfaces and from their computers. Nonetheless, deleting from these platforms requires different mental models and this can be a challenge for users as most of them assume these platforms work the same way and expect the same results. In this paper, we seek to bridge this knowledge gap.

We conducted an exploratory study using semi-structured interviews [4] to explore users' motivations for deletion, their successes/failures with regards to deletion, their coping strategies upon failures and their wants with regards to deletion in the cloud. We interviewed 26 active cloud users from a wide range of backgrounds and used a grounded theory approach [8] to analyze the insights from the interviews in order to explain why users behave the way they do, or why they make the decisions they make.

Contributions. Our analysis reveals that cloud users fail to delete in the cloud because they: (i) lack information on deletion; (ii) have incorrect mental models of deletion; or (iii) because they have to deal with poorly designed cloud interfaces. Users stated that there is not enough informa-

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

tion on deletion while information on benefits such as storage size is transparent on advertisements. We discovered that this lack of sufficient information on deletion leads to construction of inaccurate mental models which result in poor decision making and incorrect actions with regards to deletion.

We also learnt that users develop different coping strategies to circumvent their deletion challenges; these strategies are based on their needs and reasons to delete. For example, users with low deletion skills may exclude files they perceive important or confidential from the cloud so that they do not have to deal with deleting them. Surprisingly, we found that none of our participants, including those with privacy concerns, used encryption tools, although literature on cloud deletion proposes the use of encryption [25, 29].

In summary, the novel contributions of our work are as follows:

- We identify four key drivers that motivate users to delete from the cloud.
- We reveal why users fail to delete from the cloud, highlighting what causes and contributes to such failures.
- We uncover different coping strategies adopted by cloud users to address their deletion challenges, discussing the consequences of such strategies with regards to users' motivations to delete.
- We reveal what cloud users want regarding deletion from the cloud, and discuss open challenges and present paths for future research in this area.

This paper is structured as follows: Section 2 gives an overview of related work. Section 3 describes our methodological approach and demographics. Section 4 presents our findings regarding current deletion practices, challenges, coping strategies and what deletion experiences do users want. In Section 5, we discuss how coping strategies relate to users' motivations to delete and their mental models respectively. Section 6 concludes the paper.

2. RELATED WORK

Prior studies have investigated specific areas of security-related behavior like ours, but none have specifically looked at users' perception of deletion in the cloud. In non-cloud contexts such as social media, some aspects of deletion have been explored, for instance, there has been work focusing on understanding users' privacy concerns over social media and their challenges of deleting from such platforms [2, 19, 16, 27]. Some prior studies [19, 31, 21] have found that people use deletion as a coping strategy to handle regrets over posts and in some cases to protect their privacy either by removing a contact (e.g., unfriending) or the post itself. To contextualize our study, we now discuss briefly the most relevant prior work.

2.1 Users' perception of security and privacy in the cloud

Some prior work has investigated users' perception of security and privacy in cloud computing. Ion et al. examined users' attitudes and beliefs regarding the usage of cloud computing comparing them to those of enterprises [18]. They found a significant mismatch between users' security expectations and what was actually being offered. Gashami et al. researched the needs and privacy concerns of individuals adopting Software-as-a-Service (SaaS) [14]. They found that perceived benefits had a more direct effect on users' intention to adopt SaaS while privacy concerns had no direct impact. Daniel et al. investigated users' acceptance of the cloud with regards to trust and risk [5]. Their work focused on understanding how trust and risk perceptions influence users' cloud adoption decisions. Susan et al. found that, while journalists used some cloud tools, a number of them did not perceive any security risks in doing so [23]. These works have explored security as a general concern for users but have not investigated deletion as a specific problem for cloud users.

2.2 Mental models

A number of researchers have adopted the mental models approach to understand how non-expert users perceive privacy and security. Wash has examined how home computer users make their security decisions and how that affects their use of security advice [32]. He identified eight (8) mental models, "folk models", of security threats that users constructed and how these models can be used to justify why home computer users ignore security advice. Camp proposed five (5) possible models that could be used to communicate complex security to normal users [7]. Both Wash and Camp found that users want security but, in different situations, their desire to have security also depends on how they understand and perceive risk. Using a mental model approach, Bravo-Lillo et al. [3] conducted studies to gain an understanding of how users perceive and respond to computer alerts while Ur et al. examined whether users' mental models of password security matched reality [30]. Some studies have also compared how the mental models of experts and non-experts differ, for example, [17]. All these works demonstrate that understanding users' mental models can help to communicate with or educate users on security risks. Our work contributes to this area by trying to understand what deletion mental models cloud users possess.

3. METHODOLOGY

We conducted a qualitative inquiry into how cloud users understand and think about deletion in the cloud. We carried out semi-structured interviews with 26 participants between November and December 2016.

3.1 Ethical considerations

Our study was approved by the relevant Institutional review board (IRB) before any research activities began. We obtained informed written consent from all participants to take part in the study and to have the interviews audio recorded.

3.2 Participants/Sampling methodology

We recruited our participants through our existing professional networks, word of mouth and also advertised the study through posters in the city and around our institution. Interested participants were invited to complete an online form which contained a set of questions designed to screen participants that could be invited for one-to-one interviews.

Other than for balancing demographics, the screener also focused on asking participants about the cloud services they were currently using and the devices they used to connect to such services. Respondents were also asked about their activities in the cloud, that is, whether they have ever deleted, shared or uploaded and downloaded data from the cloud.

Over a period of 3 weeks, we received a total of 48 responses. From these, 16 were males, 28 were students (3 doing masters degrees while 12 were pursuing a Ph.D.), 18 had some form of employment and 2 were unemployed or retired. The majority (30) were between the ages of 21 and 30 years old.

All 48 of our respondents stated that they used smart phones to connect to the cloud while 43 used their laptops to connect to cloud services. 97% of our respondents stated having shared, uploaded and deleted data from their cloud services. Appendix A.1 summarizes the demographics of all our respondents.

All screener responses were analyzed as we tried to identify a group of about 20 to 25 participants for one-to-one interviews, a number that was enough to reach a saturation point in terms of the emerging codes as discussed in Section 3.4 . For maximum variation, we purposefully selected respondents from a wide variety of backgrounds, ages, education and socio-economic classes. More importantly, we were targeting respondents who use more than one cloud service, preferably a storage service (e.g., Dropbox) and data processing service (e.g., Google Docs). Preference was also given to those participants who had mentioned that they had uploaded, shared and deleted data from these services. Interviewing respondents from a wide range of backgrounds allowed us to capture different perceptions of deletion in the cloud and identify common patterns. In the end, 26 out of the 48 respondents to the preliminary screener were invited by email to participate in the interview. The sample included 14 women and 12 men, between 18 and 50 years old. Appendix A.2 summarizes the demographics of participants invited for interviews. For a 30 to 45 minutes interview, participants were compensated $10 for their time and effort.

3.3 Interviews

Interviews were led and conducted by one researcher in different places to meet participants' needs and requirements (e.g., at a participant workplace because of their work schedule). The vast majority of interviews (25) were conducted in-person, though a single one was conducted via Skype.

We began each interview by first obtaining consent and explaining the purpose of the study. We used a semi-structured interview protocol so that our list of questions could act as a guide throughout the interviews but not restrict us to just those set of questions [24]. Using semi-structured interviews allowed us to probe participants for more information.

We used a reflective questioning technique to interview our participants. This allowed participants to reflect their actions and decisions aloud hence not directing them to a conclusion. Reflective technique also gave us an opportunity to explore a participant's knowledge, skills, experiences, attitudes, beliefs, and values. Our questions focused on the general use of the cloud and deletion of data from the cloud. We asked participants how they use cloud services on a day-to-day basis and their reasons behind using and choosing such services. Regarding deletion, we asked participants about how and why they delete data in the cloud and the situations when they encountered problems when attempting to delete from the cloud. Some of our questions included scenarios (see below) that required participants to use their mental models to make decisions on how to delete data.

All interviews were audio recorded using a secure audio recorder and stored securely. The audio recordings and the transcriptions were not accessible to anyone other than the researchers and the transcribers.

3.3.1 Scenarios

As part of understanding users' perception of deletion, we used two scenarios and asked participants to describe what would happen when deleting under each scenario. By doing this, we gave our interviewees the opportunity to apply their mental models of the cloud and deletion. We chose these scenarios as they represent typical deletion tasks associated with cloud storage services. Each scenario was contextualized based on the information the interviewee provided and then narrated to the participant. For example, if the interviewee mentioned that they regularly used Dropbox to share photos, then the scenario would involve Dropbox and sharing of photos. We wanted to create a scenario that appeared real to the interviewee. The two scenarios are as follows:

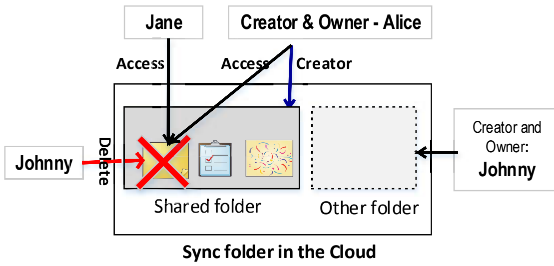
Deleting from a shared folder. This scenario (shown in Fig. 1a) asked participants what would happen if they deleted a file from a shared folder created by their colleague or friend. In this scenario, Alice has created and shared a folder with both Jane and Johnny. Johnny is running out of space but decides that the only file he can delete is the one in this shared folder because he has finished using it. However, without first contacting Jane and Alice, Johnny has to make the decision whether to delete the file or not. Would Johnny be able to delete this file? If he deletes this file, what would happen?

Deleting a shared folder. The second scenario (shown in Fig. 1b) asked users what would happen if they tried to delete a shared folder created by their colleague or friend. In this scenario, Alice has created a project folder and shared it with both Jane and Johnny. After the project has been completed, Johnny thinks he has no use for all the files in the shared folder. Johnny goes to his laptop and deletes the shared folder from his sync folder. Would Johnny be able to delete the folder? If he deletes it, what would happen? Will Jane and Alice still have access to the files in the shared folder or the folder will disappear from both of their accounts? Or will it only disappear from Johnny's sync folder?

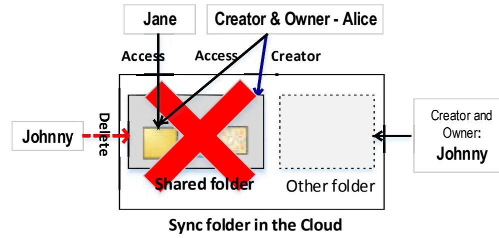
3.4 Grounded Theory

After the first five interviews, we transcribed the audio files and began coding. Beginning coding as soon as we received data was important because it allowed us to identify interesting categories and themes which needed further exploration [8, 15] during subsequent interviews. Data was analyzed through several iterative stages of open, axial and selective coding and constant comparisons of codes [8, 15]. Using NVIVO[1], we went through each transcript line-by-line and developed our first descriptive open codes. Several codes about how cloud users use the cloud and how they delete began to emerge. This process resulted in 120 unique codes. To verify the codes, after the main coder had coded the first two scripts, the second researcher independently coded two

[1]http://www.qsrinternational.com/what-is-nvivo



(a) Deleting a shared file



(b) Deleting a shared folder

Figure 1: Deletion scenarios: (a) We asked a user "Johnny" what would happen if he deleted a file shared between him and his friends "Jane and Alice" created by "Alice", and (b) we asked what would happen if Johnny attempted to delete a shared folder instead of the file.

other scripts resulting in a subset of codes from the main coder and the codebook was modified accordingly [20].

Our second phase of coding involved identifying patterns, connections, and relationships between the codes we initially developed. By doing this, we grouped similar or related codes to form categories (concepts) and in some cases expand the codes themselves to make categories. As different groups continued to emerge, we began to compare the groups against each other looking for connections between them. While additional interviews were performed, we continued with coding and memo writing until no more new codes were emerging. New codes stopped emerging after analyzing 13 interviews,that is, we reached saturation in grounded theory terms. In this phase, memos were used to describe codes, events, behaviors, and record emerging questions during the study. Following this, we ordered and further grouped our categories into more broad and abstract groupings. In the last round of data collection, we added a few more questions to the interview based on our groupings and the questions that emerged from our memos. For example, in our second phase of data analysis, two participant mentioned they would want their photos to be completely deleted, so we began to ask our next set of participants what kind of information would they want to see completely deleted.

The last phase of coding involved selective coding, where further transcripts were analyzed and we attempted to identify a linking core category that describes the underlying phenomenon in the observed and interpreted behavior. This iteration gave us the chance to engage more with the study, understanding what the users were saying and doing with regards to deletion.

4. KEY FINDINGS

Fig. 2 presents an overview of our key findings which we summarize below:

(1) What makes them delete? Our analysis suggests that users' motivation to delete falls into four major categories: privacy-driven, policy-driven, expertise-driven and storage-driven.

(2) What causes deletion failures? Not everyone can delete when they want to. Failure to delete in the cloud is not merely caused by poorly designed user interfaces but rather this can be attributed to a lot of other factors which may include inaccurate mental models and lack of sufficient information on deletion.

(3) How do users cope with deletion failures? Users develop and choose a coping strategy based on their motivation to delete or the cause of their failure to delete. For example, users whose intention to delete is privacy-driven will always choose a strategy that will remove the file from the cloud or will stop uploading the files they perceive to be important or confidential. Whereas, users whose reason to delete is to gain more storage space, will not mind cloud hopping to gain additional storage.

(4) What do users want? Users desire four key characteristics with respect to deletion in the cloud: transparency in deletion, deletion to be complete, control over deletion, and help service to support deletion tasks.

We next detail each of these findings:

4.1 What motivates users to delete?

Before we could try to understand why users could not delete, it was important to first understand what motivates users to delete or the situations in which people want to delete. Users' motivations to delete were: privacy-, expertise-, policy- or storage-driven.

4.1.1 Privacy-driven Motivations.

Users' concerns about online information, the level of trust the user has towards the provider, or the perceived negative consequences of not deleting a file from the cloud are often driving factors for deletion.

Lack of trust in provider. Users with high privacy concerns towards cloud providers are always motivated to delete. Participants revealed that they delete because they fear that their data may fall into the wrong hands. For instance, P4 said,

"*It's just me being cautious that this, from my understanding and it's not that good, Dropbox is something like an online database like or a storage space, so I prefer just to be on the safe side to delete everything so that nobody else can access to these files apart from me.*"

P4 later on continued,

"*. . . I don't want anything bad to happen or Dropbox being hacked. . . I have interviews with children, so I send recordings of interviews with children so just to be on the safe side*

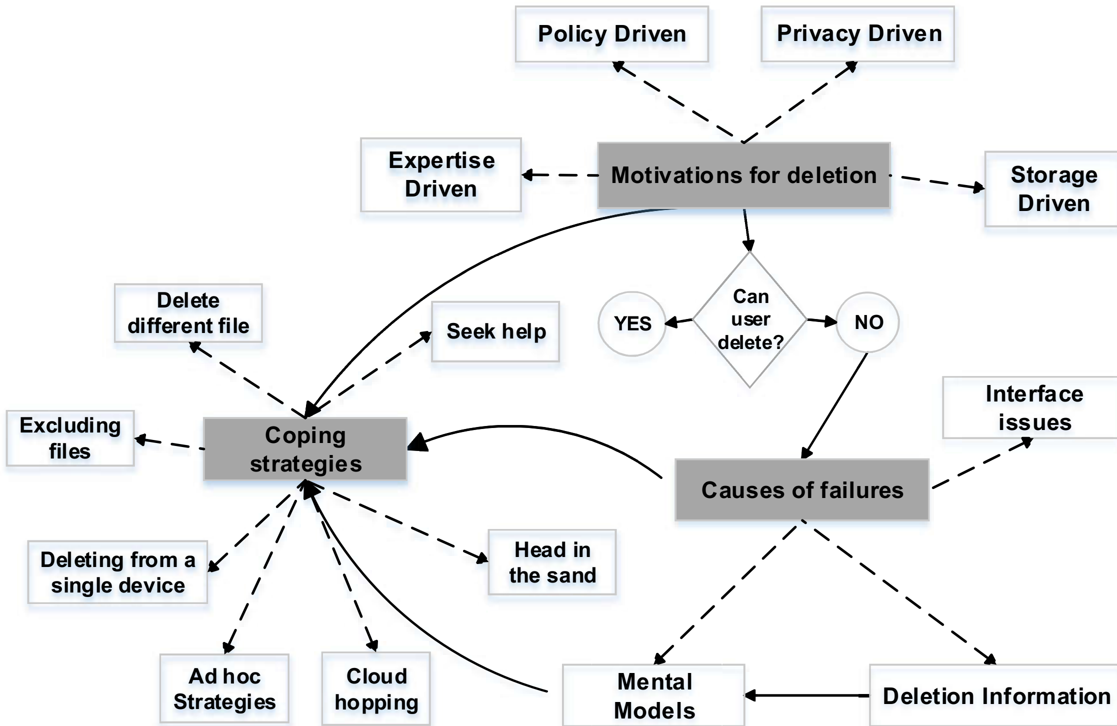


Figure 2: Key Findings

once she receives everything, I delete everything."

To avoid future conflicts. Participants deleted to avoid future conflicts which may be unearthed by the data they have in the cloud.

"*... I no longer want the pictures of my lover to be accessible to anybody else I'd want them gone from the servers forever, because my next future [partner] might discover them ...*" P11

"*I get rid of things that could come back to haunt me...*" P10

To forget. Some users delete in order to forget. Most participants who automatically save their photos in the cloud revealed to us that they commonly delete photos from the cloud in order to forget about them, e.g., photos which are perceived to be embarrassing or contain an unpleasant memory.

"*... I'm always deleting pictures and stuff that I don't need to remember.*" P20

"*Some unhappy memories maybe, I would want them [to] disappear forever and I don't want to see them again.*" P17

4.1.2 Expertise-driven Motivations.

These are factors that are motivated by the level of understanding a user has or their ability to delete successfully.

Self-efficacy. Users' desire to delete is heavily influenced by their confidence in their ability to complete the desired task. Participants who had enough knowledge or skill to delete tended to make the decision to delete whenever they wanted and execute it immediately. However, those who struggled to delete showed less interest in wanting to do so.

Deleting after unintentional use. As highlighted in other cloud studies [9], our early coding resulted in the category of "unintentional use". Several of our participants deleted from the cloud because they first used it without knowing. Nowadays, it is common to find cloud-based applications already installed in smart devices and computers. However, at first, most users are not aware that some of these cloud applications will automatically log them into these services and start saving data to the cloud. Upon realization, most users' response was to delete the data as soon as possible. Participants *rushed* to delete because they were not sure how their data got there in the first place and, because they were not sure whether their data was public and, hence, visible to everyone. One participant who did not realize she was using OneDrive for 2 years noted:

"*When I first found out I had it I tried to delete all the photos because I got scared and I managed to..., I deleted the files and I got really confused when I first opened it because I was like well how did these files end up here? I never put them there but obviously, I'd whacked them in my phone and that's where it'd automatically saved to. So I deleted them.*" P18

4.1.3 Policy-driven Motivations.

Users are driven to delete due to *extrinsic* policies, e.g., organizational security policies to which they must adhere, as well as *intrinsic* ones, e.g., perceived value of information held in a file.

Organizational policy compliance. Compliance with organizational security and information sharing policies is often a driving factor for deletion. Participants mentioned that their work policies required them to manage data securely which included deletion. However, interviewees also revealed that they continued to use public clouds – despite this being in violation of organizational policies – because they were convenient and easy to use.

"*. . . I use [Box] because it was recommended and we were told that we couldn't store research work on Dropbox . . . Sometimes I use Dropbox . . . most people use it, but if I use Dropbox I delete...*" P9

Perceived value. Users' decision to delete was also influenced by the usefulness, sensitivity and value of a file. When users perceived a file to be confidential or more sensitive or valuable, they would want to delete it immediately after using it. Also, if users considered a file to be no longer useful or needed, they would consider it a good candidate for deletion.

"*. . . It's just normal deleting. I read the paper and if it's no use for me, then I delete it, that's it. . . There was document we had, it was not good for our job, we just deleted [it].*" P8

4.1.4 Storage-driven Motivations.

Users are also driven by storage size and the need to organize their data systematically.

Storage size. The most prominent repertoire was deleting in order to free some space. Instead of buying more storage some cloud users use deletion to reclaim used space. The less space a user has, the more motivated they will be to delete.

"*I didn't have a lot of space so I had to take out some pictures and videos so I deleted them from iCloud.*" P16

"*[When my] space was limited I would actually go through and prune out what's important what's not.*" P11

Tidying up. Sometimes cloud users delete to keep things tidy. We discovered that when users have the skill and the knowledge to delete, they will sometimes take time out just to tidy things up from their cloud accounts.

4.2 Why do users fail to delete?

Although the exact details of users' failure to delete varied, our data suggests a range of common factors that lead to deletion failure. These factors include insufficient deletion information, mental models, and user interface designs.

4.2.1 Insufficient information

Insufficient information on deletion contributed to lack of understanding on deletion. Although we did not ask our participants about information on deletion, participants notified us that such information is hardly available. It also emerged that service provider advertisements had information on the benefits of using the cloud (e.g., storage size) but none on deletion.

"*Nothing like that is made very clear when you sign up. Maybe if you read through the gazillion terms and conditions you would find out, but there's nothing obvious that I've come across anyway that says, 'After this amount of period of time this will actually be deleted.' So, surely that should be one of the first things that they tell people.*" P19

Users sometimes find deletion messages difficult to understand and not providing them information that is pertinent to the deletion task.

"*Sometimes on iCloud it does not allow you to delete, like if you are trying to delete something it says that if you delete it will mess up everything else, but on Google Drive and on Dropbox, I've never found anything like that. On ICloud sometimes it does not allow you to delete some stuff.*" P24

Although such information is sometimes made available in the terms and conditions of services, our interviews show that users do not read the terms and conditions, therefore, they do not have a full understanding of how their data will be handled by the provider. Concepts such as retention period are unknown among users. Previous research on privacy policies and terms and conditions states that users perceive terms and conditions as being long and unreadable[12, 22]. Some of our participants noted that, while they did not have problems with deleting, there was insufficient information on whether their deletion is permanent or not.

4.2.2 Mental Models

Users have been known to construct their own mental models when they are insufficiently educated on an issue but have to complete a task [32]. However, most of these models are inaccurate and in most cases lead to wrong results. Unsurprisingly, most cloud users with incomplete mental models of data management in the cloud reported not being able to delete. Users' inaccurate mental models did not just lead to deletion failure but also to wrong expectations.

Our data shows that most cloud users have less or minimal understanding of the cloud and of deletion. For example, some users presented limited knowledge on how the synchronization folder works or that they can access the cloud independent of apps that consumed or used the cloud. In terms of deletion, they did not understand the concept of *Deleted folder* (i.e, where deleted data goes), retention period, and different levels of deletion. Using these themes, we identified several mental models that existed among our participants concerning deletion and the cloud. In table 1 we show which models were common among our participants and which one were not. Also, since some of our participants owned more than one model, we do not report how many participants owned each model.

Table 1: Common mental models

Popular Models	Uncommon Models
- Sync folders are not the cloud - Saving and deleting work the same way - The cloud within an app - Borrowed deletion models - Shared folder: Deletion is one sided	- Shared folder: Deletion affects all members - Providers don't delete

Sync folders are not the cloud. To automatically save data in the cloud, cloud users have to install synchronization software or applications in their devices. During installation, a synchronization application will create a local folder in the user's computer which will be linked with the cloud service.

The purpose of the synchronization application is to detect all the changes (e.g., adding or removing files) in the folder and update user's contents in the cloud. This ensures that user's files in the cloud and the local device are up to date. For this to work, a user is required to be logged in to their cloud account through the application. Although, this is popular amongst users, our data suggests that most of them do not understand how this works. Some users deleted from the sync folder while their computers were offline but expected the files to be deleted instantly from the cloud. Also, another group did not understand that deleting from this folder would also mean deletion from the cloud.

"*So once I put [my] files in Box sync folder and it uploaded automatically but I wondering if deleting my files in that folder would also delete from the cloud as well... but then I just deleted those files from my folder and then logged into my box [account] and found [that the] files in the cloud were deleted as well. I was not happy to see that.*" P24

Saving and deleting work the same way. Some users wrongly concluded that deletion in the cloud worked the same way as saving a file. Although this is correct to some extent, that is only so when using a sync folder in a computer. When using sync folder, every file placed in that folder will automatically be uploaded and saved in the cloud. When a file is deleted from a sync folder it will be removed from the user space visible to the user but placed in a *Deleted folder* in the cloud. However, this may not apply in a situation where a user connects to the cloud through another app (e.g., camera app to backup photos). Deleting a photo from the camera app may not necessarily delete it from the cloud. However, some participants expected files to be deleted from the cloud when they deleted them from their mobile devices because they automatically saved to the cloud. One frustrated participant said,

"*I used to think that it[deletion] was kind of automatic that if I deleted from the phone that it would like [delete] because the fact that I save the photos, that it's saved in iCloud I think that if I delete it ,it would delete itself from my iCloud as well.*" P14

The cloud within an app. The use of applications to consume cloud services has left many users not knowing that they can access the cloud independently from those applications. This type of users do not delete from the cloud because they do not know that their data may still be in the cloud and they can access the cloud to delete. Users also do not know that some applications may backup data automatically to the cloud. One participant was surprised when asked if she had ever directly logged in and deleted from her iCloud account:

"*I [have] never come to the conclusion that I could actually go there and delete*" P14

Borrowed deletion models. Some users have transferred their mental models of deletion from other online services such as online social networks to the cloud. When asked about how they would delete a file from a shared folder in the cloud, one Dropbox user responded saying:

"*... I think I will ask my friend to delete it. And if they don't then I can't do anything apart from untag myself. I think it's quite a similar policy like in Instagram or in Facebook when you want to delete it, it always gives you the options either contact your friend to have them delete the photo or you just unfriend them.*" P14

Shared folder: Deletion is one sided. The concept of deleting from a shared folder is a challenging and confusing one for most users. Some users believed that when one deletes from a shared folder, the deletion will only remove a file from their accounts but that particular file would still be available for other members of the shared folder. To these users, deletion from a shared folder is one-sided. We also found out that some users believed that when a file is uploaded to a shared folder in the cloud, the cloud would create multiple copies of the file for all the members of the shared folder. Thus, assuming that when they delete from a shared folder, they are only deleting their own copy and not deleting from everyone. One participant likened deleting from a shared folder to deleting from Whatsapp messenger:

"*... [If she deletes it] it wouldn't get deleted from my side but it will obviously just get deleted from her side [because] she doesn't want it. Like if you send a Whatsapp message. So normally it wouldn't get deleted from my end, it would just get deleted from her end.*" P12

Shared folder: Deletion affects all members. Some users reported that they knew that deleting from a shared folder may remove the file from all other members' accounts as well. Nevertheless, this caused a conflict within the user who no longer needs a shared file within a shared folder and wishes to delete it. Users find it difficult to make the decision whether to delete or not. They believe that there is no alternative way of deleting a file from a shared folder while other members of the folder are still in need of that file. We found that users who possess this model prefer to leave the file in the shared folder undeleted just to be on the safe side.

"*... when I do this it's always after my transcriber has used the material and sent me the transcriptions back so I always think it is safe to delete them now because in my head I'm thinking... if I delete them she won't be able to see them, so I wait for her to finish the job and then I delete them.*" P4

Deletion is permanent and instant. Some of our subjects had a very under-developed model of deletion in the cloud. These subjects did not think about deletion in any depth but concluded that it was instant and permanent. The fact that a file disappears from their sight was enough for them to conclude so. These subjects were unconcerned about their deleted data and believed they were safe after deletion. Our data further suggested that this belief affected their views on recovery from the cloud; according to these subjects data recovery from the cloud is not possible.

We discovered a conflict within the minds of some of the individuals who had this model. On one hand, they believed deletion was permanent because they could not recover deleted data in the cloud. On the other hand, they also believed that deletion in the cloud could not be permanent since cloud is an online service. This conflict was caused by the belief that online services do not delete data therefore cloud being an online service would also not delete data. A handful of participants, however, did not fall into these conflicting models. They suggested deletion in the cloud was not permanent and even constructed attack models for deletion:

"*...you delete something but they still keep a copy of it and then some can hack in and get your information. I think because they keep a copy of everything, so I think after deleting they still keep a copy,... it means that somewhere they keep information that could be retrieved later, but whether that information is kept confidentially or [if] it could be hacked, people hacking in and getting other people's pictures and then blackmailing them and stuff.*" P24

Providers don't delete. A group of participants believed cloud providers do not delete for their own benefit like advertisements and research. They held the view that there is a "secondary" storage where deleted data is stored but users are not able to access it. Although, they reported high privacy concerns, they also exhibited some defeatism:

"*I never read the T&Cs I don't really know if it's deleted forever. It's probably still stored somewhere, but I don't have immediate access to it.*" P11

"*I think the provider might not want that[deletion] to happen because they keep the data and then they want to use it for marketing and, you know, different purposes.*" P24

However, some participants were not concerned about their data being used for adverts because they believed that as long as they could achieve what they wanted to do, this was acceptable.

"*I'm not so much concerned with that [deletion] as to the underlying reasons, and the drivers for the business are more important to me than their terms and conditions.*" P26

4.3 User interface issues

As expected, we uncovered several cloud interface issues that negatively affected users' deletion process. Poorly designed user interfaces caused a lot of distress to some users which resulted in them losing interest in deleting or left them frustrated. Users are affected by screen sizes, type of interface (i.e., whether it is a web or mobile application), and how the deletion process is completed in that application. Our data also suggests that when users find it difficult to use an interface to delete, they are unlikely to attempt to delete using the same interface in the future. A number of our participants who access the cloud through mobile phones reported that sometimes they do not know where to go in their mobile phone interfaces in order to delete from the cloud.

Effort. Some cloud features (e.g., auto saving) affected users' deletion process. Users who have auto synchronization turned on ended up not deleting from the cloud because a lot of effort may be required from them to delete all the unwanted files synchronized to their cloud storage. One participant who had this feature turned on informed us that sometimes their smart phone accidentally takes photos while in their pocket leaving them with a lot of unwanted photos. It required a lot of effort and time to delete such photos from their phone and then from cloud, as a result sometimes they chose not to delete from the cloud.

"*...often with mobiles these days if the camera goes off in your pocket, which it often does, you can end up with all these blank photographs. Of course they go into OneDrive, so you look at your OneDrive and you want to clear all that. But sometimes it can take ages.*" P23

Buggy software. Buggy applications and unresponsive interfaces left some users not being able to delete. Some users reported that sometimes when they try to delete, the app or web interface would not respond resulting in them abandoning the process. Less satisfying mobile apps and unresponsive web interfaces resulted in users having less desire to delete. For example, some users reported that for them to delete they have to try it a couple of times before the operation successfully completes. Users found this annoying and preferred not to try deleting when they wanted.

4.4 Coping mechanisms

Our analysis reveals that users have developed different coping mechanisms to address or mitigate their failure to delete from the cloud. These mechanisms ranging from ignoring deletion altogether to changing cloud providers through to seeking help from others and ad-hoc strategies. We discuss various coping mechanisms employed by our participants next.

4.4.1 Head in the sand

Most participants who could not delete preferred to leave their files undeleted in the cloud. We identified four reasons why users settled on this strategy: (i) They perceive this method to be easy and quick—it does not require them to put in any effort. Users who felt deletion can be burdensome preferred this method. (ii) When a user has sufficient storage space left on their account, they are highly likely to leave the files in the cloud. (iii) When users perceive the file to be harmless or non-confidential. (iv) Trust in the cloud provider: when users trust the service provider they are more inclined to leave data in the cloud.

4.4.2 Cloud hopping

Those with high privacy concerns or low-levels of trust in providers often opted to stop using certain cloud services or changed their service providers. Although, this may be considered an extreme measure, we discovered that people weighed the benefits of using a cloud service against the cost of their undeleted data being exposed. Others noted that they changed providers because of running out of storage space with their current provider. Interestingly, we found that none of the participants who changed providers deleted or deactivated their accounts with the previous provider.

4.4.3 Excluding certain files from the cloud

Some participants reported that they explicitly decide on what goes in to the cloud before they upload to the cloud. By excluding potentially sensitive or confidential documents and sharing them by offline means, such users believed they were safe and they did not have to worry about undeleted information. This approach is common among people who have high privacy concerns and low trust towards providers. However, this strategy may open up other threats in terms of data exposure, e.g, through lost removable (potentially unencrypted) media—some participants reported using USB sticks to share sensitive files.

4.4.4 Deleting from one device

Some users reported that they only deleted from devices they were more comfortable with and were confident would yield expected results. For example, some users opt to delete from their computers (sync folder) than to delete from the web interface or mobile applications.

"*Yes, I have the app on my phone but I rarely use it, my app. I have downloaded the app on my desktop. So, I delete from there instead... I mostly delete it from the desktop because I found it difficult to delete it on the phone.*" P24

This strategy does delete the file from the cloud, however, this may lead to delays to deleting a file because the user may not always have the device they are comfortable with all the time.

4.4.5 Seeking help

Several users reported that they ask for help from their friends, family and colleagues if they think it is urgent and important that a file should be deleted. Others revealed that they will search online for solutions, for example, from tutorials, forums and blogs. The type of help sought depends on users' motivation to delete. For example, when the motivation to delete is due to high privacy concerns or trust issues, then the user will not hesitate to ask their social network for help. However, when the file to be deleted is sensitive (e.g., explicit photos) or confidential, users have a likelihood of not seeking help from other individuals with the fear of being exposed or shamed. They would opt for looking for solutions online. This strategy leads to high chances of deleting from the cloud and participants who reported they knew how to delete revealed this is how they learnt about deletion.

4.4.6 Deleting a different file

Some of our participants reported that when they are deleting to free up space and they fail to delete a certain file, they will instead choose a different file to delete than the original one. One participant explained that sometimes they get a warning not to delete a file but because they do not fully understand the consequences of deleting that file, they will instead look for a different file to delete.

" *...I will not delete that one, I will try to find something else to delete instead of it, to get more space otherwise I can't. So, at times I will remove a different file because I have not [yet] found a correct solution on how to get over it.*" P24

4.4.7 Ad hoc strategies

We also discovered some ad hoc strategies among our participants. Some of them revealed that they did not have a well-defined method of overcoming the failure to delete. They reported that they will try to find the best possible solution that fits and suitable for that moment in time. Some participants reported that if they cannot delete a file, but they feel it is important to delete such a file from the cloud, they would try deleting it from all their devices including the web interface. One participant revealed that when they are in need of more storage and they cannot delete they will simply buy more storage.

4.5 What deletion experience do users want?

In the previous section we discussed different coping mechanisms employed by users when they struggle to delete. Here we focus on themes and categories that emerged as to what cloud users want. We identified four key themes across our participants.

4.5.1 Transparency in deletion

Participants want providers to be more transparent about deletion of data; they wanted information on how their data is deleted to be made freely available. Participants who struggled to delete suggested that providers should provide tutorials on how to delete, and that deletion information should be made clear when they first sign up for services.

However, those who could delete were not so much interested in such information but rather in knowing more about how deletion is done in the cloud. Some suggested that they should receive notifications when data is completely removed from the cloud. That is they want guarantees that their data is completely gone from the provider. We found that this suggestion was popular among participants with privacy-driven deletion practices.

4.5.2 Complete deletion

Early in our research process, the assumption that cloud deletion was complete and instant emerged and we formulated some interview questions around this belief. As a result, we asked users what deletion meant to them and most of them defined it as "getting rid of" or "destroying data". Our analysis suggests that they believe deletion in the cloud is complete because they use these mental models when they think of the cloud. However, we found that participants who had better knowledge of the cloud and deletion desired complete deletion. They suggested that deletion should mean permanent:

"*I suppose all data should be completely deleted. Once you press 'delete', delete should mean delete, so then you don't have that sort of grey area as to what's sensitive, what's not sensitive. Delete should mean delete, I think.*" P6

"*The moment I delete something from my iCloud, or my laptop, I want it to be deleted completely. I feel like once I [have] said I don't want this on my laptop again, or I don't want it on my phone, I would rather have it deleted everywhere, complete.*" P3

For users who assumed deletion was complete, we explained to them that it is possible that their data may not be completely removed from the cloud [26]. Most of our participants responded with shock to this revelation while others reported they had always thought it might not be deleted.

"*I've always had it in the back of my mind that what you delete does not completely go. I didn't know like it's almost impossible to delete...*" P3

Although most users reported they would want complete deletion, this wasn't true universally. A number of users, understandably, wanted to have the opportunity to recover deleted data especially data deleted by mistake.

With regards to complete deletion, several of our participants exhibited the following beliefs: (1) data perceived important or confidential should be completely deleted, (2) Data belonging to other people or data that contains identifiable information should be completely deleted, and (3) only law abiding citizens should be allowed to completely delete things from the cloud.

With regards to the final point, we discovered that users who perceived themselves as harmless and law abiding citizens did not mind if their data was not completely deleted. However, they reported that if the data did not belong to them or contained other people's identifiable information then they would want it gone. Some users believed that complete deletion would enable law breaking citizens to commit and hide their crimes on the cloud. Despite this, other users reasoned that they would still choose permanent deletion because it is their data and no one has the right to access it after deletion. With regards to recovery, such users reported they would change the way they work and just be careful when deleting data. We found that participants who belong to this group were people who had high privacy concerns about the cloud and would rather lose data because of mistakes than have it undeleted in the cloud.

4.5.3 Contact Point

During the interviews, some users reported that it was hard for them to get verifiable information on using the cloud. They suggested that a service dedicated to resolving their cloud queries would be useful especially when they cannot ask anyone. One participant put it this way:

"*[First thing is,] I don't know whom to call. If I want my data back I don't know whom to call, whom to contact because I don't think they have any helpline or service like this where a customer can call and say, 'I deleted by mistake, send it me back,' or I don't know whether the provider has the access to retrieve particular data of a customer.*" P24

4.5.4 Control over deletion

Our analysis suggests that users feel the need control over deletion in the cloud. They wanted to be the ones who decide when a file should be permanently deleted or held for potential recovery. Our data also shows that users want control over what is synced over to the cloud so that they will not have to delete.

"*Because if I have deleted something, I am saying, 'I don't need it anymore,' or, 'I don't want evidence of it anymore,' then surely it should be deleted completely because I no longer have use for it. Who's meant to still have use of what I've deleted?*" P22

5. DISCUSSION

Table 2: A summary of motivations to delete and coping strategies

Motivation to delete	Coping Strategy
Privacy-driven	Seeking help
	Deleting from single device
	Excluding certain files from the cloud
	Cloud hopping
	Ad hoc strategies
Expertise-driven	Head in the sand
	Ad hoc strategies
	Seeking help
	Excluding certain files from the cloud
	Cloud hopping
Policy-driven	Ad hoc strategies
	Deleting from single device
	Seeking help
	Excluding certain files from the cloud
Storage-driven	Cloud hopping
	Delete a different file
	Deleting from a single device
	Ad hoc strategies

5.1 Deletion motivations and coping strategies

Our findings reveal that users' choices and development of coping strategies are dependent on their motivation to delete. These relationships are summarized in Table 2 and discussed next.

Motivation: Privacy-driven. Users whose motivation to delete is privacy-driven are always quick to seek help in deleting or have a higher chance of employing some ad hoc methods to try to delete from the cloud. Seeking help from other cloud users has a likelihood of deleting a file from the cloud. Ad hoc strategies do not always guarantee data will be deleted from the cloud. When struggling to delete some of these users may opt to delete from the device they are most comfortable with or confident that they will manage to delete data using it. This choice is normally based on users' past experience; the user chooses it because it has worked for them before. Users who try all the above strategies and still fail, are inclined to change their provider or decide not to ever upload files they perceive to be confidential. These strategies are perceived to provide maximum privacy by the user as sensitive files would never reach the cloud from which they struggle to delete.

Motivation: Expertise-driven. Expertise-driven users often resort to ad hoc strategies when they cannot delete. If ad hoc strategies do not work, those with high self-confidence would either decide to leave the file in the cloud or hop to another provider. Such users do not normally ask for help because of their self-belief. Users with less skill and low confidence in using the cloud are likely to leave the file undeleted but not change the provider. They will only change the provider if they are confident of using the platforms/interfaces from the new provider, because they do not want to go through the process of having to relearn how to use the new cloud provider. Expertise-driven users may also resort to excluding sensitive files when using the cloud to avoid the anxiety associated with not being able to delete the file from the cloud.

Motivation: Policy-driven. Users whose motivation to delete is policy-driven usually fear the consequences of having that data not deleted in the public cloud. They usually adopt ad hoc strategies as their first coping mechanism, and if they still cannot delete from the cloud, they would then attempt to delete from the device they are confident in using. They would finally ask for help if everything they tried has failed. Nonetheless, prior failure to get data deleted causes them to exclude valuable or work-related files from the cloud.

Motivation: Storage-driven. Users who delete for storage reasons adopt strategies such as deleting a different file, deleting from a single device, cloud hopping and some ad hoc strategies. Deleting a different file may temporarily create space, but as the number of files that the user cannot delete increases, the user eventually runs out of space. Deleting from a single device and some ad hoc strategies may yield results since files get deleted. However, ad hoc strategies like buying more storage cost the user but do not lead to fulfilling the initial goal – that of deleting from the cloud. The results of cloud hopping are temporary; it only works until the user fills out all the new storage provided. In general, a cloud hopping strategy does not scale as users are unlikely to keep changing providers regularly.

5.2 Mental models and coping strategies

We also observed a potential connection between users' mental models and coping strategies. In Table 3 we summarize how users' mental models and coping strategies are linked.

Table 3: A summary of users' mental models and their coping strategies

Mental models	Coping Strategy
The cloud within an app	Seeking help Deleting from single device Excluding certain files from the cloud Cloud hopping Ad hoc strategies Head in the sand
Borrowed mental models	Head in the sand Ad hoc strategies Seeking help Cloud hopping
Sync folders are not the cloud	Ad hoc strategies Deleting from single device Seeking help
Providers don't delete	Cloud hopping Excluding certain files from the cloud
Shared folder: Deletion is one sided	Head in the sand
Shared folder: Deletion affects all members	Head in the sand Seeking help
Deleting and saving work the same way	Seeking help Cloud hopping
Deletion is permanent and instant	Head in the sand

The cloud within an app. Users who believe the cloud is inaccessible are likely to seek help in order to delete since they do not believe they can actually log on and delete. Others may just try to delete using the single device that they believe is connected to the cloud, which may not be effective as discussed earlier. Some users may choose to cloud hop in search for a cloud that they believe they can access and delete data from, or they may leave files undeleted. Some privacy-driven users with this mental model may opt to exclude files from the cloud, only storing files they perceive to be not confidential.

Borrowed mental models. Upon realizing that their deletion understanding is different from cloud deletion, users may choose head in the sand approach, change to a new provider, seek help, or try other ad hoc strategies. They may choose head in the sand approach because they recognize the mismatch and unexpected outcome. Such users may hop to another cloud where such mental models may yield expected results, and they may finally seek help after trying some ad hoc strategies.

Sync folders are not the cloud. Users who believe sync folders are not part of the cloud may adopt ad hoc strategies to delete from the cloud. Some may seek help to delete while others may delete from web interface instead of the sync folder.

Providers don't delete. These users are likely to change providers or choose to exclude files they perceive to be important from the cloud. Although changing a provider does not solve their deletion issues, they believe the new provider will provide a certain assurance of deletion.

Shared folder: deletion models. Interestingly, though shared folder mental models are different, in both cases, users employ the same coping strategy, head in the sand. Users who perceive deletion to be one sided may choose to leave the files undeleted believing that deleting would not delete the files from other members of the shared folder, while, those who perceive that deletion affects all the members of the shared folder may choose to ignore the file because they do not want to delete files which other users are still using. Users who believe that deletion affects all members of the shared folder often seek help to confirm whether they could delete from the shared folder.

Deleting is permanent and instant. The main challenge faced by users with this model is decision making when it comes to deciding whether a file should be deleted or not. They will believe that it will get deleted forever and instantly, as a result, these users often resort to leaving files undeleted in the cloud with a belief that they might need them again.

Deleting and saving work the same way. Users who possess this model tend to seek help when encountering problems with deletion. Some may also choose to delete from the device where this mental model accurately applies, hence leading to successful outcomes. In this case users rely on recalling previous successful deletion experiences.

5.3 Design implications

Our results revealed a major gap in users' understanding of how the cloud and deletion work. Although the responsibility to delete lies between the providers and the users, our study pointed out that cloud users want more transparency regarding cloud deletion policies. Information on deletion should be clear and easily made available for users especially about how data is disposed after use. Users would also benefit from cloud providers making deletion mechanisms easy to understand and accessible.

Regarding help, cloud users would like to have the option to contact someone directly concerning their deletion problems. This implies that some cloud interfaces provide users with not enough feedback or complicated information which is hard to understand. Something akin to Deletion Service Points would help users resolve their problems quickly. With regards to user interfaces, improving user feedback (e.g., notifications during deletion) would inform users on the end results of their actions therefore influencing or improving their weak mental models.

Another possible avenue for improving the current situation is improving users' understanding and awareness-building. Our study found that users possess different mental models at the same time, of which most are incomplete and lead to failure to delete. We argue that these differences make user education a challenging task hence such education should be customized. Also, since not all incomplete mental models lead to failure to delete, we suggest that user education should focus on maturing mental models that are weak or those that lead to failure to delete.

5.4 Limitations

Our study is a qualitative inquiry – based on a sample size of 26 participants. This sample is significant for such a study and saturation in grounded theory was reached by 13 transcripts. As such we can be confident that the motivations, failures, coping strategies and desires discussed in Section 4 are grounded in the data from the study. We also accounted for coding bias by using a second researcher to verify the codes emerging from the grounded theory analysis. However, some of our participants who reported that they used the cloud for editing documents could not distinguish between Microsoft Office Online, Office 365 and Office 2016. This may have influenced their judgment and perception of deletion from the cloud. At the same time, it further reflects the inaccuracy of users' mental models with regards to the cloud. Future studies ought to explore the users' mental models of the cloud in general and their impact on various user interactions with the cloud with regards to security and privacy.

5.5 Further research

Our study can form the basis of a number of research directions that can contribute to better understanding and supporting users' needs with regards to deletion in the cloud and associated security and privacy goals.

Deep understanding of cloud users' mental models

In this study, we uncovered different mental models constructed mainly by those participants who could not delete, and we observed that these mental models seem to have an influence on users' deletion practices and behaviors. Hence, we realize the importance of understanding other deletion models constructed by users, and also mental models about the cloud in general, particularly focusing on whether users use or transfer these models to the cloud from other domains (computers, smartphones, etc.) or whether they develop new ones to cope with a new reality. This is important in order to understand the extent of the influence of mental models regarding cloud users' practices and behaviors, as well as their adoption of the cloud. It would also be interesting to study security experts about their cloud usage with respect to deletion. Understanding how they use and delete from the cloud could shed light on the differences between them and the findings of this paper.

Understanding the impact of users' practices and behaviors

In our study, we found that people were using the public cloud for their work purposes without any security tools such as encryption. However, we are yet not sure what motivates this behavior and whether users are aware of the risks associated with this behavior, and how do they decide what goes into a public cloud and what does not. Future research should explore this issue in order to develop further insights into how these behaviors affect users' privacy and security, and the privacy and security of their organizations.

Cloud deletion in specific domains

Cloud usage and deletion could also be explored considering different types of organizations, such as governmental organizations, and across different type of industry organizations. It would be interesting to explore if deletion strategies, failures and coping mechanisms are domain-dependent.

Evaluation of encryption tools and deletion

Some studies (e.g., Tang et al. [29], Rahumed et al. [25] and Ramokapane et al. [26]) recommend the use of encryption tools in the cloud to protect users' privacy after deletion. However, none of our participants mentioned the use of encryption as a means to assured deletion. It is not clear whether users are not using such mechanisms because of a lack of awareness or due to usability issues. Usability studies in this area would help understand how such tools could be improved, or how users could be encouraged to adopt them.

Multiparty access control

In our study, we revealed that cloud users possess incomplete mental models about deleting from *shared* folders, which are managed by one or more users. Even if these models were complete and accurate, the issue of data management, and in particular data deletion, when multiple users are involved in the cloud is an under-explored area. Such multi-party access control has been studied in other domains such as social networks [28] and it would be interesting to study the applicability and usability of such techniques in order to support deletion in the cloud.

Follow-up confirmatory studies

Since our study was of an exploratory nature, we identified factors that play a role in deletion in the cloud and potential relationships between them grounded in the data obtained through semi-structured interviews. The next step would be to undertake confirmatory studies, to further understand these concepts and confirm the extent of their relationships.

6. CONCLUSIONS

Although it is generally assumed that deletion is an easy task, our study shows that cloud users struggle to delete. Their failure to delete leads to construction or development of coping mechanisms to address the problem. Users develop these strategies if they believe that it is important that data is deleted from the cloud. However, information on deletion affects how users construct deletion mental models. A lack of information on deletion leads to construction of incomplete or inaccurate mental models which eventually leads to a failure to delete. These mental models have a direct impact on the choices and the development of coping strategies. Incomplete or inaccurate mental models may lead to development of strategies which do not delete data from the cloud, or strategies which only solve the problem temporarily or bring up additional problems. All in all, our investigations bear out the intuition that usability of deletion or lack thereof in the cloud is a key privacy and security challenge that needs significant attention.

7. ACKNOWLEDGMENTS

We thank all the people who took part in our study. We also thank Dr. Asad Naqvi for his assistance in coding and analyzing our data, and all the anonymous reviewers for their helpful comments and suggestions.

8. REFERENCES

[1] A. Adams and M. A. Sasse. Users are not the enemy. *Communications of the ACM*, 42(12):40–46, 1999.

[2] H. Almuhimedi, S. Wilson, B. Liu, N. Sadeh, and A. Acquisti. Tweets are forever: a large-scale quantitative analysis of deleted tweets. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 897–908. ACM, 2013.

[3] 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, 2011.

[4] A. Bryman. *Social research methods*. Oxford university press, 2015.

[5] D. Burda and F. Teuteberg. The role of trust and risk perceptions in cloud archivingâĂŤresults from an empirical study. *The Journal of High Technology Management Research*, 25(2):172–187, 2014.

[6] C. Cachin, K. Haralambiev, H.-C. Hsiao, and A. Sorniotti. Policy-based secure deletion. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 259–270. ACM, 2013.

[7] L. J. Camp. Mental models of privacy and security. *IEEE Technology and society magazine*, 28(3), 2009.

[8] K. Charmaz. *Constructing grounded theory*. Sage, 2014.

[9] J. W. Clark, P. Snyder, D. McCoy, and C. Kanich. I saw images i didn't even know i had: Understanding user perceptions of cloud storage privacy. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1641–1644. ACM, 2015.

[10] S. Clark, T. Goodspeed, P. Metzger, Z. Wasserman, K. Xu, and M. Blaze. Why (special agent) johnny (still) can't encrypt: A security analysis of the APCO project 25 two-way radio system. In *20th USENIX Security Symposium, San Francisco, CA, USA, August 8-12, 2011, Proceedings*, 2011.

[11] B. Computer. *Dropbox Deleted files*, 2017 (accessed February, 27 2017). `https://www.bleepingcomputer.com/news/software/dropbox-kept-files-around-for-years-due-to-delete-bug/`.

[12] L. F. Cranor. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *J. on Telecomm. & High Tech. L.*, 10:273, 2012.

[13] Forbes.com. *Apple safari web history*, 2017 (accessed February 28, 2017). `https://www.forbes.com/sites/thomasbrewster/2017/02/09/apple-safari-web-history-deleted-stored-icloud/#7e58cad76328`.

[14] J. P. G. Gashami, Y. Chang, J. J. Rho, and M.-C. Park. Privacy concerns and benefits in saas adoption by individual users: A trade-off approach. *Information Development*, 32(4):837–852, 2016.

[15] B. G. Glaser and A. L. Strauss. *The discovery of grounded theory: Strategies for qualitative research*. Transaction publishers, 2009.

[16] P. Ilia, I. Polakis, E. Athanasopoulos, F. Maggi, and S. Ioannidis. Face/off: Preventing privacy leakage from photos in social networks. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 781–792. ACM, 2015.

[17] I. Ion, R. Reeder, and S. Consolvo. ... no one can hack my mind": Comparing expert and non-expert security practices. In *Proc. SOUPS*, 2015.

[18] I. Ion, N. Sachdeva, P. Kumaraguru, and S. Čapkun. Home is safer than the cloud!: privacy concerns for consumer cloud storage. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, page 13. ACM, 2011.

[19] M. Johnson, S. Egelman, and S. M. Bellovin. Facebook and privacy: it's complicated. In *Proceedings of the eighth symposium on usable privacy and security*, page 9. ACM, 2012.

[20] K. M. MacQueen, E. McLellan, K. Kay, and B. Milstein. Codebook development for team-based qualitative analysis. *CAM Journal*, 10(2):31–36, 1998.

[21] A. E. Marwick et al. Social privacy in networked publics: TeensâĂŹ attitudes, practices, and strategies. 2011.

[22] A. M. McDonald and L. F. Cranor. The cost of reading privacy policies. *ISJLP*, 4:543, 2008.

[23] S. E. McGregor, P. Charters, T. Holliday, and F. Roesner. Investigating the computer security practices and needs of journalists. In *USENIX Security*, pages 399–414, 2015.

[24] S. Portigal. *Interviewing users*. Rosenfeld Media, 2013.

[25] A. Rahumed, H. C. Chen, Y. Tang, P. P. Lee, and J. C. Lui. A secure cloud backup system with assured deletion and version control. In *Parallel Processing Workshops (ICPPW), 2011 40th International Conference on*, pages 160–167. IEEE, 2011.

[26] K. M. Ramokapane, A. Rashid, and J. M. Such. Assured deletion in the cloud: requirements, challenges and future directions. In *Proceedings of the 2016 ACM on Cloud Computing Security Workshop*, pages 97–108. ACM, 2016.

[27] M. Sleeper, J. Cranshaw, P. G. Kelley, B. Ur, A. Acquisti, L. F. Cranor, and N. Sadeh. I read my twitter the next morning and was astonished: A conversational perspective on twitter regrets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3277–3286. ACM, 2013.

[28] J. M. Such and N. Criado. Resolving Multi-Party Privacy Conflicts in Social Media. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1851–1863, 2016.

[29] Y. Tang, P. P. Lee, J. C. Lui, and R. Perlman. Fade: Secure overlay cloud storage with file assured deletion. In *International Conference on Security and Privacy in Communication Systems*, pages 380–397. Springer, 2010.

[30] B. Ur, J. Bees, S. M. Segreti, L. Bauer, N. Christin, and L. F. Cranor. Do users' perceptions of password security match reality? In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3748–3760. ACM, 2016.

[31] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, and L. F. Cranor. I regretted the minute i

pressed share: A qualitative study of regrets on facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*, page 10. ACM, 2011.

[32] R. Wash. Folk models of home computer security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, page 11. ACM, 2010.

[33] A. Whitten and J. D. Tygar. Why johnny can't encrypt: A usability evaluation of pgp 5.0. In *Usenix Security*, volume 1999, 1999.

APPENDIX

A. DEMOGRAPHICS

A.1 Respondents Demographics

Table 4: Summary: Respondents Demographics.
A total of 48 people responded to our advert, the table below summarizes their demographics.

	No. of participants
Gender	
Male	16
Female	32
Age	
18 - 20	10
21 - 25	18
26 - 30	11
31 - 35	5
36 - 40	2
41 - 45	0
46 - 50	1
51 - 55	1
Educational Background	
High school/College course	14
Bachelors	13
Masters	11
PhD	9
Preferred not to say	1
Employment status	
Unemployed/Retired	2
Full time	15
Part-time	3
Student	28
Cloud services	
Dropbox	22
iCloud	20
OneDrive	22
Google Drive	21
Box	13
Microsoft Office 365	35
Google Docs	29
OneNote	6
Cloud Access	
Smartphone	48
Tablet	24
Desktop	30
Laptop	43
Cloud activities	
Uploaded files	48
Deleted data	47
Shared folder/files	46
Deleted an account	30
Recovered deleted files	15
Downloaded files	36
Read a service agreement	12
None of the above	1

A.2 Interview Demographics

Table 5: Summary: Interview Demographics.
26 people were invited to take part in our interviews.

	No. of participants
Gender	
Male	12
Female	14
Age	
18 - 20	3
21 - 25	8
26 - 30	6
31 - 40	7
45 +	2
Educational Background	
High school/College course	5
Bachelors	9
Masters	5
PhD	6
Preferred not to say	1
Employment status	
Unemployed/Retired	1
Full time	12
Part-time	3
Student	10
Cloud services	
Dropbox	15
iCloud	9
OneDrive	6
Google Drive	17
Box	17
Microsoft Office 365	15
Google Docs	14
OneNote	2
Cloud Access	
Smartphone	26
Tablet	13
Desktop	23
Laptop	25

B. INTERVIEW GUIDE

Thank you for participating in our study. As you read in the consent form, we will be recording the session so we can review it to make sure that we don't miss any part of our conversation. Your information will be kept confidential and will only be accessed by us. Your name will not be associated with any data I collect. Do you have any questions regarding the consent form? Do I have your permission to start the recording?

1. Do you use any of the following services or similar services? Examples: Dropbox, Box, iCloud, G-Drive, One-drive.
   - Follow-up-1: How often do you use them?
     (a) Prompt: Would you say you use them every day?

- Follow-up-2:What do you use these services for?
  (a) Prompt: Is it for work or its personal?
- Follow-up-3: You mentioned that you use [service/services], how do you use [it/them].

2. Do you use any of the following services or related services? Examples: Office365, Google Docs etc.
   - Follow-up-1: How often do you use them?
   - Follow-up-2: What do you use these services for?
     (a) Prompt: Is it for work or its personal?
   - Follow-up-3: Can you describe to me how you use [name of the service]?
3. Do you have any particular reason why you use these services?
4. When you store your files in [service mentioned in Q1] or create a document in [service mentioned in Q2] what happens?
   - Prompt: Do you know where they are stored?
5. Have you ever deleted something you uploaded on [service mentioned in question 1]?
   - Prompt: Have you ever thought of deleting something you have uploaded online?
   - Follow-up-1: Why?
   - Follow-up-2: Can you share with me how you go about deleting a file in [service mentioned by user in Q1]?
     (a) Prompt: You mentioned that you use [name of the service], how do you delete data from [name of the service]?
6. Have you ever deleted something you uploaded on [service mentioned in question 2]?
   - Follow-up-1: Why?
   - Follow-up-2: Can you share with me how you go about deleting a file in [service mentioned by user in Q2]?
     (a) Prompt: You mentioned that you use [name of the service], how do you delete data from [name of the service]?

   [NOTE: If the participant claims to have never deleted anything from the cloud before, ask the following question otherwise skip it]
7. You have mentioned that you have never deleted or been asked to delete anything before, how come?
   - Follow-up-1: How do you deal with information that you no longer need?
8. Have you ever faced problems or challenges when trying to delete your data from any of your services?
   - Prompt: Can you recall a time when you wanted to delete something but could not figure out how to delete it or you could simply not just delete.

   [If participant says Yes]
   - Follow-up-1: Which service was that and how did you resolve or get around those challenges? Or how did you finally delete then?
9. Have you ever been required to recover information you have previously deleted?
   - Prompt: Have you ever needed a document or file that you had previously deleted from [service mentioned in Q1 or Q2].
   - Follow-up-1: Were you successful?
   - Follow-up-2: How did you do it?
10. Do you ever think the information [e.g., files, documents] you have previously deleted still exist somewhere online or can be shared by your service provider?

    [If participant says Yes]
    - Follow-up-1: Why?
    - Follow-up-2: What do you do to ensure that your deleted information can never be shared after you have deleted it?

    [If participant says No]
    - Follow-up-1: You mentioned that you don't think your information could be shared after it has been deleted, why?
11. After you delete your files, do you know how long it takes for [service mentioned at Q1 or Q2] to delete them from their side?
    - Prompt: How long does deletion process take?

[Explain to the participant that you will share a scenario with them and then ask questions using the scenario. Choose one scenario per interview depending on the interviewee occupation, for example, if the interviewee is a student ask them scenario one.]

Scenario 1
After a [late night out/party/picnic], your [friend/colleague] creates a folder in [service mentioned] and shares it with you and your other friend. He then tries to be funny and decides to upload 3 embarrassing photos of you three that you took on the night out. You are embarrassed and decide to delete all the photos from the shared folder.

Scenario 2
You have just joined a new team at work. Your new supervisor creates a folder in [service mentioned by participant] and shares it with you and your other colleagues. Your supervisor uploads some documents for you and your team to work on. Upon a discussion between you and your supervisor, s/he realizes you don't need one of the documents so s/he asks you to delete the document.

[Scenario Questions]

12. What do you think will happen when you delete the [photos/document]?
    - Prompt: Will [they/it] be deleted from the shared folder or just your computer or device?
13. Will the [photos/document] be deleted from all your [friends'/colleagues'] accounts or they will only be deleted from your account?
    - Prompt: Will the deletion process affect your [friends/colleagues] too?

[End of scenario questions]
Explain to the interviewee that the questions on the scenario have ended.

14. If you were told that information you delete may never be completely deleted, what would you do differently?
15. Do you know anything about the "right to be forgotten" European ruling?

[Explain to the user that you are at the end of the interview, ask them if they do have any questions or anything they want to share about deletion from the cloud.]

The Importance of Visibility for Folk Theories of Sensor Data

Emilee Rader
Department of Media and Information
Michigan State University
emilee@msu.edu

Janine Slaker
Department of Media and Information
Michigan State University
slakerja@msu.edu

ABSTRACT

Sensor-enabled wearable devices and smartphones collect data about users' movements, location, and private spaces and activities. As with many ubiquitous computing technologies, this data collection happens in the background and appears "seamless" or invisible to the user. Despite this, users are still expected to make informed choices regarding consent to data collection. Folk theories are sets of beliefs and understandings that arise informally and guide decision-making. To investigate folk theories regarding sensor data collection that might guide privacy self-management decisions, we conducted qualitative free list activities with 30 activity tracker users in which we asked them to list "information that an activity tracker knows". We found that folk theories regarding the data that activity trackers collect depend on interactions between the users and their trackers that provide visibility into dependencies among data types, evidence about what trackers are able to record, and feedback that inspires speculation about how trackers work. Our findings suggest opportunities for designing interfaces that intentionally support the development of folk theories about how sensor data are produced and how they might be used, which may enable users to make more informed privacy self-management decisions.

1. INTRODUCTION

Ubiquitous computing systems that incorporate a wide variety of sensor technologies are an increasingly common part of everyday life for many people. In particular, wearable devices like smart watches and fitness bands, and smartphones carried in a pocket or purse, have been widely adopted. All of these devices include embedded sensors that engage in continuous data collection, and are capable of producing inferences that users consider "extremely private" [39]. For example, in February 2016 a Reddit user posted heart rate data from his wife's Fitbit activity tracker to enlist the community's help in troubleshooting what he believed was a malfunctioning device. Instead, he found out from other users that what he had noticed could actually be valid data indicating that his wife might be pregnant (in fact, she was) [17].

User awareness and concern regarding data sharing and use often receive more attention in the security and privacy literature than data collection does. Privacy concern has been shown to depend on contextual aspects of sharing and use [30, 33], and encouraging people to think about different possible audiences and uses can affect how concerned they are [23]. However, data collection practices have also long been recognized as related to privacy. This was acknowledged in the original 1973 report on which the Fair Information Practice Principles (FIPPs) were based [52][1]. It was again recently emphasized in the 2012 "Consumer Privacy Bill of Rights" report issued by the Obama administration [47], which includes the directive, "Consumers have a right to exercise control over what personal data companies collect from them and how they use it."

People are expected to be able to self-manage their privacy by making decisions about what systems to use and what kinds of data collection to consent to [45]. This approach assumes that all users are able to think and behave correctly, and in an informed and rational fashion, which is not realistic [22]. An approach adopted by security and privacy researchers regarding how to understand users' choices and behavior focuses on folk theories related to technology use [2, 53, 57]. Folk theories are beliefs, analogies and explanations that guide people's behavior, which develop and evolve through everyday experiences. Folk theories about how technologies work form even when details about the inner workings of the technologies are invisible to users [11, 37]. By investigating folk theories related to sensor data collection, we can gain insights into how everyday interactions with sensor-enabled systems support their formation. We can also find out more about what guides users' privacy self-management decisions and behavior regarding these systems.

We conducted a qualitative study focusing on folk theories about data collected by activity trackers, defined as smartphone apps and standalone devices that support fitness-related data collection (movement, heart rate, etc.). These devices and smartphone apps are an example of sensor-enabled technologies that have achieved wide, mainstream use. They also have an interface that provides information to users about the data they collect; seeing step counts and other health and fitness activity information is part of the reason why people use them. Our focus is considerably more narrow than studies like Wash's folk models of threats [53] and Yao et al.'s folk models of online behavioral advertising [57], and more like Kempton's study of thermostats which focused on a single application [22].

We found that participants' folk theories conceptualized types of data their trackers were collecting as if they were either manually *entered* by the user, directly *measured* by the tracker, or *calculated*

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

[1] "There must be a way for an individual to find out what information about him is in a record and how it is used." [52]

from other data the tracker had collected. Participants' folk theories had developed through interactions with their trackers that provided visibility into dependencies between different kinds of data, such as values presented in the interface that increase at the same time, visible heart rate sensors, and step counts shown by the tracker which did not seem to match participants' perceptions of their activity. However, participants' folk theories had not incorporated the idea that their data were estimates or inferences. This precluded speculation and reasoning about how raw sensor data might be useful for other purposes outside the context of activity tracking that could reveal information about activities and personal characteristics users might not want to disclose. In other words, these folk theories would not support users being able to consent to uses beyond the context of activity tracking.

With this paper, we make several contributions. We present findings about folk theories related to sensor data collection in ubiquitous computing systems that provide insight into what users are aware of and can reason about, in their own words and from their own perspective. We highlight the importance of making aspects of the data visible in the interface, and of designs that encourage users to speculate about the origins of their data. And, we present design implications for ways activity tracker interfaces might better support speculation, and thereby the formation of folk theories that help users reason and make decisions about privacy.

2. RELATED WORK

2.1 Folk Theories

Folk theories are "ways of understanding" that help people interpret phenomena they encounter in the world [12]. They are based on experience and interaction rather than formal instruction, and are often shared across people [21, 22]. Folk theories for specific technologies arise out of users' everyday experiences with those technologies [4, 22]. This means they sometimes vary from person to person, and are often incomplete and inaccurate from an expert point of view [11, 4]. Folk theories help users generate explanations [18], guide inferences they make about cause and effect [4], help them reason about what technologies are capable of [35], and influence their choices and decision-making with respect to those technologies [53]. Folk theories are also sometimes called folk models, or mental models, and are elicited through qualitative investigations that involve interviews and hypothetical scenarios [53, 56], activities like sketching [19, 57], and prompts such as photos [35] or specific app permissions [24] that participants react to. Many studies have found that folk theories held by experts are different from those held by non-experts; for examples, see Asgharpour et al. [2], Kang et al. [19], and Yao et al. [57].

In a widely-cited study, Wash [53] investigated folk theories that non-expert computer users have for security threats like hackers and viruses, and argued that they have implications for whether people believe particular pieces of security advice are important to follow or not. Kang et al. [19] investigated folk theories of the internet, and found that non-experts drew simpler diagrams of locations where data is stored online and where it travels to, whereas experts drew more complex diagrams with more parts and components. Both experts and non-experts knew that their data is shared with companies that provide services to them. But beyond that, all participants expressed a lot of uncertainty. A more recent study about online behavioral advertising found that folk theories involve beliefs about agency, including which entities are involved in tracking users, and where the data are stored. These ranged from a browser-pull model where the browser is responsible for storing all user data and obtaining relevant ads, to a more technically accurate model that involved both first- and third-parties [57]. Folk theories about RFID, which is sensor-related technology, have been found to be partially correct as well: the most common folk theory involved the idea that RFID tags are small devices that can hold a little bit of information, similar to a magnetic strip or a barcode [35]. And, related to activity trackers in particular, Yang et al. [56] found that many activity tracker users had engaged in "ad-hoc assessment" of how their trackers worked while they used them, which resulted in feelings of frustration related to perceived inaccuracies in their activity tracker data. Yang et al. recommended that users should be provided guidelines for how to determine the accuracy of their devices, calibration mechanisms so trackers can better adjust to individual variation, and training resources that explain what trackers measure in order to increase understanding. These studies are all examples of investigations into folk theories of technology use, and serve as background about the approach we take in our study, and the kinds of insights about users, and about design, that this approach supports.

2.2 Wearables, Smartphones and Privacy

A classic early paper about wearable sensor technology and privacy describes the Active Badge system [51]. This system was developed in the early 1990s as a proof of concept location tracking system for a research organization. It used RFID technology to track users' locations via wearable badges that could be detected by a sensor array in the building. Many people found the system to be useful, in that it enabled them to find colleagues more easily when they wanted to talk to them. However, people also expressed concerns about privacy, mostly surrounding ways that the location data provided by wearing the badge could potentially be used, and by whom. This theme about data use comes up in many studies about sensor data and privacy. Even when users report that they feel the collection of the sensor data itself does not concern them, when asked to consider possible uses they are able to imagine harms that might result if the data were used improperly.

For example, Klasnja et al. [23] asked participants about privacy concerns related to their use of the UbiFit wearable fitness prototype over a 3-month period. None of the participants were concerned about the idea of data collection, because they didn't think the data were sensitive on their own. However, they had concerns about the use of location data, and raw audio data. Similarly, Raij et al. [38] showed participants who used the AutoSense system for three days visualizations of the data that had been collected about them. Participants reported increasing concern about more sensitive kinds of data (e.g., less concerned about physical activity, more concerned about conversations and stress level). In contrast, Rapp et al. [39] and Motti and Caine [29] both found that users of commercial activity trackers did consider the data that was collected to be extremely private. But, in both studies participants remarked about concern due to a feeling that they were not necessarily in control of how their data might be shared with third parties and subsqeuently used.

Other researchers have focused on the issue of smartphone apps accessing and potentially sharing user data with third parties, or data "leakage". Shklovski et al. [44] interviewed smartphone users about this, and found that it is a source of concern for users. Balebako et al. [3] took this a step further by notifying participants when smartphone apps were accessing data; participants were surprised about how often this happened, and more concerned about it than they were before they were made aware of it. Almuhimedi et al. [1] also focused on smartphone notifications in their research, and used the notifications to provide a means of awareness and control over smartphone app data access and use. In their study, the notifications showed users information about how many apps had accessed different types of smartphone data in a specified time period. This style of intervention assumes that drawing users' attention to data sharing and use by informing users about how many times different apps were accessing their information would raise concern and "nudge" users to take action. Just over half of their participants in this study made changes to app permissions as a result of the intervention.

Shih et al. [42] also studied smartphone users' privacy concerns and willingness to share data, via a custom app they created. The app was designed to measure participants' privacy preferences regarding app usage of personal data by asking them questions periodically over the 4-week period of the study involving different combinations of app types, data types, and usage purposes. Users were least willing to share information when more details were given to them like the name of the app that was using the information, and what the app was going to do with the information. In other words, providing more detail about use was associated with less willingness to share the information.

We focus in this paper on an application of sensor data collection that is commercially available and in mainstream use: wearable activity trackers and smartphones used for the purpose of activity tracking. This sampling frame enabled us to recruit participants who had already been using sensor-enabled devices and smartphones for their own reasons, some of them for a number of years. Activity trackers already support some form of user interaction with the data they collect, which presents researchers with an opportunity to study folk theories of sensor data collection that have developed over time in actual use, rather than as a short-term research intervention.

Unlike previous work on folk theories of technology related to security and privacy, with the notable exception of the study by Poole et al. about RFID [35], our study specifically involves sensor technologies. It is also different from work focusing on privacy awareness and concern related to data sharing and use, because it focuses primarily on data collection. And, it is different from many security and privacy studies in that we do not assume there is an objectively "correct" behavior that users must be measured against. Rather, we focus on understanding non-expert users' existing folk theories from their perspective, so that we can better understand what guides their behavior, and make recommendations for design to support the formation of more privacy-relevant folk theories.

3. METHOD

3.1 Data Collection

We conducted semi-structured interviews that began with a free list activity designed to elicit folk theories about what types of information activity trackers are able to collect, and about how they are able to collect that information. Interviews lasted roughly 60 minutes and took place primarily over the phone, with a few in person, during December 2015 through February 2016. There are several advantages to conducting phone interviews versus in-person interviews. Phone interviews allow access to participants in diverse geographical locations, maintain interviewee anonymity, and can decrease social pressure and increase rapport. Research on the two methods has not found either to produce data of compromised quality [31, 46].

The free list activity lasted about 12-15 minutes, and in every case took place at the beginning of the interview, right after obtaining consent. The remainder of each interview after the free list activity focused on participant thoughts and reactions regarding a series of hypothetical scenarios in which activity tracker data might used to infer other kinds of information about the user. Each participant received a $25 Amazon.com gift card as a thank-you for participating. This study was approved by our institution's IRB. In this paper we focus on just analysis and findings from the free list part of each interview.

Free listing is a method used by anthropologists to elicit concepts that are part of a semantic domain for a group of people. Free list activities begin by the interviewer prompting the participant to "list all the kinds of X [the domain] you can think of" [6]. The interviewer then follows up with additional prompts to clarify things the participant has said and elicit additional concepts until the participant runs out of concepts to list. The goal of free listing is to gather data about the structure of a semantic domain and the relationships between concepts within the domain, as understood by the participants [55, 48]. In other words, the intent is to understand the semantic domain from the participants' perspective, not to impart any external structure onto what participants have said. Items or concepts that are mentioned infrequently, or not at all, are not considered to be part of the semantic domain according to participants [36].

Free list activities are unlike other semi-structured interview techniques in that they elicit information about things which at least "in principle have a right answer which is universally true". Participants in a free list activity should feel like they are discussing facts about the world, "perceptions, not preferences" [7]. This is an important distinction for our study, because folk theories arise out of everyday experiences in the world [22]. Therefore, we used a method to elicit participants' knowledge and understanding of the world within the semantic domain of interest to our investigation, not their attitudes, opinions, or concerns.

The wording of the domain-specific prompt we used for our free list activity was *information that an activity tracker knows*. This prompt was specifically designed to elicit concepts related to the data activity trackers collect, without priming participants to use "data"-centric terminology or focus their attention other aspects of activity trackers and data collection introduced by the researchers. The prompt did not ask participants to speculate about what might be possible for activity trackers to infer about users, instruct participants to imagine things an activity tracker *might* know, or list information that other people (instead of a device or system) might be able to infer based on activity tracker data. We avoided prompts that might encourage participants to speculate, because this could prime them to think about something they had not considered before. We wanted to elicit their existing folk theories rather than encourage them to develop new ones.

Free list activities often produce information that is incomplete or ambiguous, because recalling all associations is difficult for participants to do [8]. Most of our participants began by listing concepts related to their knowledge of activity trackers in general, and as the activity progressed they made more specific references to the tracker that they personally used. We did not direct them to focus on specific features or technical capabilities of their own particular activity tracker; rather, the prompt was intentionally general to allow participants to describe using their own language what they understood about the information that activity trackers collect. After each participant finished making his or her initial list, the interviewer read the list aloud which helped the participant to generate items they had initially forgotten to include [40]. Additional follow-up prompts were used to clarify what the participant meant

ID	Age	Gender	Activity Tracker
P01	44	F	Fitbit Flex
P02	27	F	Polar Beat App w/heart rate band
P03	32	F	Fitbit Flex
P04	48	F	Fitbit (wristband)
P05	34	F	Fitbit (unspecified)
P06	39	F	iOS Health, Move Apps
P07	30	F	Pacer App
P08	42	F	Virgin HealthMiles Pedometer
P09	32	F	Fitbit Charge HR
P10	38	F	Fitbit Charge HR
P11	23	F	Fitbit Charge HR
P12	39	F	Fitbit One
P13	40	F	Samsung S Health App
P14	24	M	Fitbit Flex
P15	36	F	LG Health App
P16	29	F	Fitbit Charge HR
P17	24	F	Google Fit App
P18	25	F	iOS Health, MyFitnessPal, WeChat Apps
P19	25	M	Argus App
P20	35	F	Fitbit Charge HR
P21	40	F	Fitbit Charge HR
P22	32	F	Samsung S Health App
P23*	34	F	Fitbit Charge HR
P24*	24	F	Fitbit Flex
P25	34	F	Fitbit (unspecified)
P26*	28	M	Fitbit One, heart rate band
P27*	33	M	iOS Health App
P28	24	F	NexTrack App
P29	36	M	iOS Health App
P30	25	M	Fitbit (unspecified)

Table 1: Participant characteristics. ID numbers with an asterisk (*) indicate participants who were no longer using an activity tracker at the time of the interview.

by the terms they listed (e.g., "What do you mean by X?" where X was the term mentioned by the participant). After the free list activity was complete, the interviewer asked the participant additional follow-up questions about the items they had listed, to elicit associations between different terms participants mentioned, and between the terms and tracker-related activities and use. For example, a follow-up question frequently asked was, "Can you tell me how you think it knows X?" (e.g., can you tell me how you think it knows steps?). The "how" prompts allowed us to elicit participants' understanding about dependencies and causal relationships between different types of information collected by their activity trackers. In the follow-up prompts, the interviewer took care to refer to concepts introduced by participants using the same terminology that they had used.

3.2 Participants

We recruited participants who were current or former users of activity trackers, which we described in our recruiting materials as *wearable activity trackers and mobile sensors that automatically count steps, like Fitbit or the Moves app*. We included both wearable devices and smartphone apps in our sampling frame because they are used for similar purposes (e.g., step counting) and collect similar data (e.g., via accelerometers). We advertised our study using snowball sampling on Facebook and email sent to a paid research pool organized by our institution. The paid research pool at that time consisted of about 3700 active users from the local community surrounding a large midwestern university. We combined these two methods of recruiting to obtain a more diverse sample in terms of geographic location [25] and demographic characteristics [41]. Roughly 60% percent of our sample came from snowballing. Friends and family members of the researchers were ineligible to participate, as were undergraduate students, and anyone who reported on the screening questionnaire that they had received training or worked as a computer programmer, software engineer, or in some other IT-related position. Folk theories of a variety of technologies have been shown in previous research to differ between experts and non-experts [2, 19, 26, 34]; the folk theories of experts are more complex and use more specialized vocabulary. We excluded technology experts from our sampling frame because we expected that they would be more familiar with how the underlying technologies work. Also, expert users may view privacy self-management differently than non-experts do. We also chose not to recruit from enthusiast venues like Quantified Self forums or to target early adopters, because we did not want to bias our sample towards self-tracking experts who might be more knowledgeable about how sensor data are produced.

Our sample consisted of 30 participants (80% female; mean age = 32.5; age range = 23–48) who lived in areas across the U.S. (e.g., Illinois, California, New York) in both urban and suburban settings. Many were administrative assistants, homemakers, and worked in research-related professions (lab managers, analysts). We also interviewed participants who worked in healthcare, state services, law and business development. Market research shows that women are more likely to own an activity tracker than men [16], and also more likely to volunteer to participate in research when online recruiting methods are used [14, 32]. While our sample was primarily female, we actively looked for evidence of differences between men and women in our data, and did not find any. All participants were current or former users of activity trackers. Eleven participants had been using a tracker for one or more years; six for 6-12 months, and four for 1-5 months. Nine participants did not report how long they had used an activity tracker. Twenty to thirty participants is a reasonable sample size for free list activities that involve a small or well-defined semantic domain, according to Weller and Romney [55]. Table 1 presents a summary of some of the characteristics of our participants and the trackers they used.

More than half of our participants used wearable devices created by Fitbit, which monitor activities ranging from step counts to sleep patterns, and provide additional information about users' activities such as active minutes and calories burned. The Fitbit Charge HR (the most popular among our sample) is distinct because it continuously monitors a user's active heart rate. Only two of the 19 participants who used a dedicated activity tracker device did not use a Fitbit. Eleven participants used an activity tracker app on their mobile phones, without a separate wearable device. These apps use sensors within the phones to track steps and other data. The Samsung S Health app uses a similar technique as the Fitbit Charge HR for measuring heart rate in which the user places her finger onto an optical sensor (located besides the phone's flash) and LED light is reflected onto the skin to determine the rate of expansion and contraction of the user's capillaries. We consider both dedicated activity tracker devices and smartphone apps to be "activity trackers" for the purpose of this study, because our participants self-identified them as activity trackers, and because according to our participants both perform similar functions and collect similar kinds of data.

3.3 Analysis

Interviews were digitally recorded and transcribed, and the transcripts were divided into two files for analysis: one containing just the initial free list activity and another for the remainder of the semi-structured interview. We analyzed the free list transcripts using an iterative, inductive coding approach which identifies themes

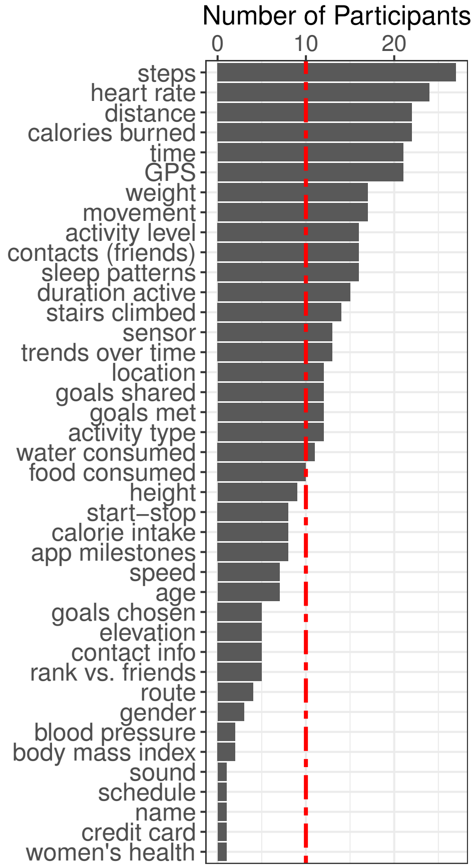


Figure 1: Histogram of the number of participants who mentioned each data type. The bars reaching beyond the dotted line are data types that at least one-third of participants mentioned.

to address "relationships of similarity" [27]. We began by standardizing phrases or statements made by participants about the same types of data so that we could generate counts of how many participants mentioned each data type [55], as is typical for analyzing free list data. While participants generally used similar words to refer to similar data types, there were some differences. For example, most participants talked about "movement", but we also coded terms like "vibration" (P16) and "jarring sensation" (P27) as movement. Participants mentioned 40 different data types; the mean number of data types they mentioned was 14.2 (*SD* = 3.86). The most commonly mentioned data type was steps taken (27 out of 30 participants); however, only 17 participants mentioned movement, which all activity trackers record to some extent. Figure 1 shows a histogram of all of the data types participants mentioned.

In addition to coding for data types, we coded for statements that participants made about *how* activity trackers come to know the information. We standardized the verbs participants used to talk about how the tracker knows, so that we could determine how many participants used these concepts and analyze which verbs were used in conjunction with the types of data that were mentioned. For example, both P14 and P22 talked about how their trackers know when they are engaged in a higher level of activity. The italicized sections of the quotes below indicate the connection each participant is making between the verb and data type they mentioned:

> If I'm moving frequently for 20 to 25, 30 minutes, I think *that gets tallied in the active minutes section*. (P14)

> I mostly run on the treadmill. So when I use the running in my app, it's not literally track[ing] the GPS so it seems like I'm not running at all because it tracks with the GPS. And so, this cannot be taken as moving, so *it's not counted as moving*. (P22)

The verbs in both of these examples, "tallying" and "counting," were standardized as "counting". Overall, participants used 32 different standardized verbs to describe how the tracker knows different types of information, and the most commonly used verb was "tracking" (15 participants), followed by "inputting" (11 participants).

After the final data collection and coding, we constructed two summary matrices [28], one consisting of transcript excerpts containing co-occurrences between different data types, and the other consisting of co-occurrences between data types and descriptions of data provenance. The matrices included only data types that were mentioned by at least 10 participants. We identified the data type(s) in each excerpt, any relationships between the data types (e.g., one information type being based on or calculated from another), and descriptions of data provenance. We used this rich dataset to generate visualizations of the connections and dependencies between data types, and identify higher-level patterns.

3.4 Limitations

The method and sampling frame we used have several limitations. We had a small sample that was selected for diversity, not generalizability. This means that our findings cannot be interpreted as statements about prevalence in a wider population of activity tracker users. Also, our qualitative data come from retrospective self reports. This is appropriate for the free list technique, but it means that we did not observe participants interacting with their activity trackers, or directly study the formation of folk theories as it happened. In addition, the data were collected by eliciting responses to a specific prompt we designed for the free list activity. There may be salient data types that participants did not mention due to the wording of the prompt and follow up questions. In particular, the choice to use a general prompt, and not to direct participants to speculate, means that we can't draw conclusions about folk theories for what activity trackers might be able to infer. Finally, because we did not ask about privacy concern as part of the free list activity, we can't use these data to connect the folk theories to specific concerns about privacy related to sensor data.

4. FINDINGS

4.1 Folk Theories about Types of Data

Our participants' folk theories about sensor data collected by activity trackers included three different categories of data types, differentiated by how they believed their trackers were able to collect or generate the data. These categories were not always technically correct compared with how activity tracker technology is actually able to generate the information provided to users, based on user documentation available from activity tracker companies and whitepapers about sensor technologies[2]. We first discuss relationships and dependencies participants described between the types of data they mentioned, and then use the pattern of dependencies to

[2]Fitbit, in particular, has extensive user documentation available on its website, help.fitbit.com, accessed on June 10, 2017.

Figure 2: The relationships between the data types mentioned by participants. Arrow thickness indicates the number of participants who mentioned a connection between a pair of data types. *Entered* data types are blue, *measured* are green, and *calculated* are red.

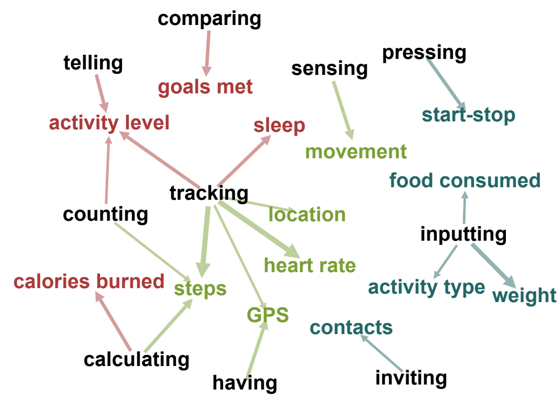


Figure 3: The relationships between data types and the verbs participants used to describe how the tracker records each data type. Arrow thickness indicates the number of participants who mentioned a connection between a pair of data types. *Entered* data types are blue, *measured* are green, and *calculated* are red.

illustrate the three higher-level categories of data that were present in participants' folk theories.

4.1.1 Dependencies Between Data Types

We identified dependencies between data types by focusing on statements participants made that indicated one data type was based on on another, for each data type mentioned by 10 or more participants. All data types participants mentioned are included in Figure 1; we focus here on dependencies between data types listed in that figure having bars beyond the dotted line, from "steps" to "food consumed". For example, P17 explained the relationships between the data types that allow her tracker to determine calories burned (emphasis added):

> I think, based on the metrics I've given it: My *age, height, weight*, so it knows all that and then it calculates *based on* my average *activity level*, how many *calories I've burned* for the day.

After identifying pairs of data types mentioned by each participant and the direction of the dependency between them, we created a network graph to visualize these relationships. This graph is presented in Figure 2. Arrows point from the *antecedent* data type (e.g., weight) to the *descendant* data type (e.g., calories burned). Only those pairwise relationships mentioned by at least three participants are included in the graph. Thicker arrows indicate that more participants talked about the existence of that relationship. Common relationships included movement to steps, GPS to location, and sensor to heart rate. Not all participants mentioned the same relationships between pairs of data types. For example, 9 participants said that steps were based on movement; however, 18 participants mentioned steps alone, without another data type.

When participants described how an activity tracker might know a certain type of information or described the relationship between a pair of data types, they often used verbs to describe the nature of the relationship. We created a second visualization (Figure 3) depicting co-occurrence between the data types and the verbs they mentioned. Arrows point from the verb to the data type, and the thickness of the arrows represents how many participants used a particular verb. For example, the verb "inputting" was used to describe how the tracker knows the user's weight. The verb "tracking" was used in conjunction with 6 different data types (steps, GPS, heart rate, location, sleep, activity level), indicating that participants used this verb in a nonspecific way to refer to something the activity tracker did to collect data. Some verbs were not used consistently across participants. "Counting" was used to talk about both steps (counting the total number of steps) and the user's activity level (what counts as an active minute). And, "calculating" was used to talk about both steps (calculating the number of steps for the day) and calories burned (calculating how many calories burned). Participants also talked about trackers "having" GPS, and "sensing" movement. Diagramming these verbs and dependencies allows us to analyze the cause and effect relationships participants described, and determine similarities between the data types participants discussed. This reveals characteristics of their folk theories.

4.1.2 Three Categories of Data

We identified three categories of data types our participants discussed: *entered, measured, and calculated* data. These categories emerged from the dependencies between data types mentioned by our participants, and the different verbs participants used to connect the data types. Table 2 shows all types of data that our participants identified during the free list activity, separated by category. Broadly speaking, *entered* data consists of data types that users manually input into their activity tracker interfaces. *Measured* data types are those participants described as being directly recorded, standalone phenomena. *Calculated* data types are considered to be generated based on other data types. These three categories are important for understanding participants' beliefs about the kinds of data collection their activity trackers engage in. We describe each of the three categories below.

Entered Data: Twenty-eight of our 30 participants reported that an activity tracker collects some information that users manually input, by entering it into a standalone tracker app or an app associated with their tracker device. All entered data types had zero antecedents (see the blue data types in Figure 2 that have no incoming arrows), and participants talked about them using verbs that indicated manual data entry. The entered data types include physical characteristics like age, gender, weight and height.

One entered data type that was mentioned by 8 participants, "start-stop", was not a characteristic of the user like weight or gender. Instead, it marked a transition between one activity state and another, such as being awake versus asleep. Participants talked about "pressing" a button (P12, P15, P24, P25), or "setting" (P05, P29), "turning on" (P05, P28) or "telling" (P01, P05) the tracker to enter sleep mode or to start or stop tracking an activity like a walk or a run. Only one participant mentioned entering information that was not directly related to fitness tracking, like name and credit card information. This is surprising, because activity tracker apps require users to create an account in order to use the service, and ask for information like name, contact information, and in some cases payment information as well.

Measured Data: All 30 of our participants believed that some data types, like heart rate, movement, steps and GPS, are direct measurements collected by the tracker. Measured data types are similar to each other in that participants mentioned virtually no antecedent data types in connection with them (incoming arrows in Figure 2; measured data types are green), or referred to them as being automatically detected by the tracker.

There are three measured data types in Figure 2, time, sensor, and GPS, that have no incoming arrows, indicating that participants believed these were not based on any other data types. However, there was some disagreement among participants about whether steps in particular had an antecedent data type. P16 provides a typical example of this:

> And yes, it definitely guesses by vibration as well, or by something like that, so it knows how many steps I take per day.

In this and similar examples (e.g., P19 said his tracker "record[s] leg movement"), participants were aware that steps are calculated based on movement. However, across our participants, steps were more similar to the other measured data types; almost twice as many participants talked about it that way. Seventeen participants did not mention an antecedent data type at all for steps, while the other 9 participants who mentioned steps said it was based on movement.

Calculated Data: Participants described calculated data types differently from entered and measured data types; all were described as having two or more antecedent data types. For example, in Figure 2, there are arrows from four different data types pointing to calories burned. P23 showed a more sophisticated understanding of this than most of our participants:

> So steps would then translate into miles traveled or some do specifically steps calculated and then they don't know, but they approximate calories burned, based on who knows what algorithm.

P23's perception that an algorithm is involved is actually fairly technically accurate. For example, the Fitbit help pages say that calories burned is "estimated based on the physical data you entered when you set up your account: gender, age, height, and weight," and "the activity recorded by your tracker"[3].

[3] `https://help.fitbit.com/articles/en_US/Help_article/1381` (accessed on July 10, 2017)

Measured		Entered		Calculated	
steps	27	weight	17	calories burned	22
heart rate	24	contacts	16	distance	22
GPS	21	activity type	12	activity level	16
time	21	goals shared	12	sleep patterns	16
movement	17	water consumed	11	duration active	15
sensor	13	food consumed	10	stairs climbed	14
location	12	height	9	trends	13
elevation	5	calorie intake	8	goals met	12
sound	1	start-stop	8	milestones	8
		age	7	speed	7
		contact info	5	friend rank	5
		goals chosen	5	route	4
		gender	3	body mass idx	2
		blood pressure	2	daily schedule	1
		credit card	1		
		name	1		
		women's health	1		

Table 2: Number of participants (out of 30) who mentioned each category and type of data. The categories and data types emerged from the data, and were identified by the researchers as described in section 4.1.

The four most common calculated data types were distance, calories burned, activity level and sleep patterns. Participants mentioned examples of calculated data much more frequently than measured or entered data, and were aware that these data resulted from combinations of other data collected by the tracker. Fifteen out of our 30 participants used the phrase "based on" to refer to the antecedents of calculated data types.

These three categories of data types highlight ways that our participants explained and reasoned about the data collection capabilities of their activity trackers. Some of these ways of understanding how their trackers work can be learned through direct experience entering personal information. However, when it comes to measured and calculated data, our participants could only rely on indirect experiences, because they were unable to observe the more technical aspects of how the trackers collect data. For example, whereas most of our participants believed that heart rate is directly measured, in reality it is inferred using optical heart rate sensors. These sensors use LED light reflected from the skin to detect changes in blood volume as a user's capillaries expand and contract. Likewise, steps are estimated using data from accelerometers, which record force exerted by acceleration, usually in three directions simultaneously. Both of these data types rely on algorithms to identify patterns in the raw sensor data to separate out the signal from the noise and estimate when a heart beat or step took place. To do this accurately requires aggregated training data collected from many different people over time, doing different activities. However, our participants' folk theories included only a highly simplified representation of this process. In the next section we provide further explanation about how visibility is important for the development of folk theories of sensor data.

4.2 Data Visibility Supports Folk Theories

Visibility into the origins of different data types was related to the categories of beliefs about entered, measured, and calculated data. Our participants' perceptions of dependencies between data types relied on evidence that they could see and verify directly for themselves; for example, by watching step counts increase while out walking. An activity tracker's user interface itself makes some data dependencies visible, like the connection between increased step counts and calories burned or active minutes. The things that participants directly and indirectly perceived influenced their reasoning about how activity trackers collect data. This section presents findings about ways that interacting with an activity tracker device and app provides experiences that help users form folk theories about what the tracker knows, and how it works.

4.2.1 Seeing Simultaneous Changes

As users see their step counts increase at the same time as other data values in the tracker's interface, they learn which data types are related to each other. For example, steps and calories burned increase together throughout the day; therefore, each step must be connected with a certain number of calories burned. P07 described it this way: "Really like however you walked, it shows how many calories that you have burned." Steps were also spoken about as a unit of measure for other calculated data types, like distance and active minutes. When P13 was asked, "you mentioned distance, can you explain what kind of distance you mean?" she answered, "Well, it goes in steps". Three other participants also talked about how steps are somehow converted into miles for the purpose of recording distance traveled (P04, P08, P23).

Many activity trackers highlight so-called 'active minutes' in the interface in order to provide feedback to users about how active they are. P14 contrasted being "constantly on the move" with "just walking around my kitchen or something" and said that his tracker can tell the difference. If he "walk[s] quickly to and from a meeting" those minutes "[get] tallied in the active minutes section" by the tracker.

Many participants talked about ways that the tracker device or app provided feedback about how much progress they had made toward the goals they had chosen. For example, P04, P23, and P30 said that their trackers vibrate when a goal is reached. Three participants talked about how the app provided feedback in the form of a visualization. P14 described how the interface looks as it changes over the course of the day as he accumulates steps, to display progress toward his step count goal:

> And obviously once it gets to the end, I know I've had my 10,000 [steps]. And the color kind of changes, so it starts off as a dark blue and then it goes to a yellow and then a green kind of once you're really approaching your step goal.

Notifications about goals met and indicators in the user interface provide information that our participants noted as signaling cause-and-effect relationships between the data types: calories burned increases because steps are increasing, vibration was caused by meeting a goal. These relationships are what participants described to us when talking about how their trackers are able to know certain types of data.

4.2.2 Direct and Indirect Perception

Characteristics of the activity being tracked assisted users in making connections between their own experience and the data collected about the activity. This is particularly clear when contrasting steps with sleep patterns. Steps were described as "counted" or "calculated", while sleep patterns were "monitored" or "watched". Steps are a discrete action from the user's perspective that are directly experienced [13]; sleep patterns occur when the user is not awake and cannot compare what the tracker says with their own perceptions. In other words, sleep was an activity "monitored" by the tracker that participants found difficult to verify, because they could not observe the quality of their own sleep for themselves.

Participants talked about both sleep patterns and steps as based on movement. However, most participants were more skeptical about what their trackers reported about sleep patterns than steps. For example, P01 told us that she does not know how the tracker can tell she is in REM sleep, saying she doesn't know much about the "inner workings" of the tracker, but that sleep tracking involves "more variables than what you can sense on your wrist". P09 said that her "completely unscientific theory" is that the tracker can tell how well she slept because it can tell when her arm stops moving and when her heart rate drops. She was satisfied with this explanation because, "it has never accidentally thought I was sleeping". In other words, she was confident in the behavior of the tracker while she was asleep, because the experience she had with it when she was awake led her to believe that it was working properly. However, P05 had the opposite experience. She had observed her tracker making the error that P09 said had never happened to her:

> But a lot of times it's not necessarily registering I'm awake, when I have a kid whose head is right on top of me and I refuse to move or something, and I'm just laying there for an hour. Like, it's not... So, I think it's doing movements.

There were in fact two participants who talked about tracking sleep patterns more confidently. These participants had less sophisticated trackers that required them to manually start sleep mode each night if they wanted to track sleep patterns. For example, P24 said,

> When I go to sleep, I have to double tap it and then it records how many hours I sleep, and it also records my movement in my sleep, so that it shows me when I'm in deep sleep because I won't be moving, and when I'm restless throughout the night it shows on a little graph.

These examples illustrate that participants were most confident about the aspects of sleep tracking that happened while they were still awake and could connect their perceptions to what the tracker was reporting, and the least confident about the aspects that they could not observe while they were asleep. Being able to connect the state transition from tracking "movement" to tracking "sleep" with something they could perceive directly, witness, or enact themselves (like seeing it change when they were very still, or manually turning on sleep mode) made sleep tracking seem more believable.

4.2.3 Visibility of the Sensor

In talking about heart rate data collection, participants were focused on the sensor—where it is worn, what it is doing, and how one can see the sensor readings. Heart rate was more strongly connected with the concept of a sensor than any of the other measured data types. Having one's heart rate measured during a visit to the doctor is a common experience, and heart rate has a medical and fitness interpretation that many people are already aware of. Many participants referred to the physical part of the body where a tracker with a heart rate sensor should be worn, which, as P09 described, is worn "a little bit above your wrist, and it has a little sensor, it's like a green light actually". P16 explained how her tracker is able to know her heart rate:

> The one that I have, the Charge HR, it measures your heart rate based on your wrist. There's a sensor that I

don't know the specifics of, that you wear one finger away from your wrist. So it's tracking your heart rate there.

Twenty four participants mentioned heart rate, but only nine participants said they used a tracker with a heart rate sensor. For example, P04 used a Fitbit wristband that does not have heart rate capability, but was aware that some trackers can do this:

> But I do know, on the wrist, that some of them can track your heart rate because obviously, that's where... If you're going to the doctor's office, that's where they're checking your pulse at.

Heart rate and movement data present an interesting point of comparison regarding participants' folk theories. Far fewer participants mentioned movement than heart rate (17 versus 24). Heart rate sensors are visible components of the device, because they must be on the outside of the device to work properly. In contrast, accelerometers, which collect movement data, are inside the device and cannot be seen. If movement was mentioned, it typically only came up as a way of explaining how the tracker was able to detect other data types like sleep patterns, flights of stairs or number of steps. For example, P27 said he was not sure about how the tracker could tell how many flights of stairs he has gone up, but said there's a "motion sensor for kind of the jarring sensation that would be given by going a single step". P09 also mentioned the sensor in relation to flights of stairs, and how she was unsure what kind of sensor allowed the device to have this capability: "And so it's obviously some kind of sensor that's just not in the other equipment [her previous tracker]."

It is as though these participants only knew about the "motion" sensor because they were trying to reverse-engineer where the calculated data values came from. However, being able to actually see a part of the heart rate sensor component (e.g., the green LED on the Fitbit Charge HR), or having to intentionally interact with it to take a measurement (put a finger on the flash, as P22 talked about with her Samsung Galaxy smartphone) makes the sensor itself more salient, making the data generated by the sensor more salient as well. This difference in visibility of the sensor and the perceptions about data provenance that visibility enabled was an important differentiator for our participants between measured and calculated data types.

4.2.4 *Perceived Inaccuracy*

Seventeen participants described noticing that their tracker counted an activity differently than they expected. For example, P23 noticed a discrepancy while applauding at a show:

> I was at a show and I clapped and I saw that [the tracker] was lighting up and then a friend of mine, who I was there with, he had a fancier, I have just the one that has lights, but his tracker actually, you could press the button and see how many steps it was. And so then after the next song we clapped again, we looked before we clapped the number of steps he had and then he clapped, and then he looked again and it was higher.

This anecdote illustrates an observation made by other researchers [13, 43, 56] who have written about the experimentation that activity tracker users engage in when they notice perceived inaccuracies in their data.

The physical display on the wrist-worn Fitbit trackers is limited, and can only display one piece of information at a time, typically a count of a data type like steps, stairs climbed or calories burned. People who wear trackers on their wrists, rather than in their pockets or elsewhere, have more opportunities to notice the disparity between their perceptions of movement and the tracker's step counts. This is because the tracker's display is more visible when worn on the wrist. These participants talked about how noticing this disagreement inspired them to assess the accuracy of their tracker's performance, and to speculate about how the calculated data are produced. For example, P03 said that when she pushes the stroller, she thinks the tracker is not "calculating" because she's "not getting any steps." P20 also made a very similar comment: she said that her tracker underestimates her activity when she is pushing the stroller or holding the dog's leash, because she isn't swinging her arm back and forth as much when doing those activities.

Perceived inaccuracies made visible by the device's display encouraged speculations about discrepancies between how the tracker works and participants' subjective perceptions of their movements. Experiencing these discrepancies provided opportunities for our participants to incorporate additional information into their folk theories about how their trackers collect data.

4.2.5 *Manual "Recording"*

Participants described using an input mechanism provided by the tracker to enter information about the beginning and end of periods of time taken up by certain kinds of activities, such as exercise or sleep. By entering this data, users can mark a change from one state of activity to another. These state changes indicating when activities start and end add a layer of context to a particular timeframe, in which the tracker then uses the sensor data it collects to determine active minutes or sleep rather than steps.

For example, many trackers offer users the ability to manually log duration and type of activity. Similar to the automatically detected active minutes, entering this information changes how the device interprets data recorded during that time period from inactive, to active. P15 talked about pressing the "record" button to enter a mode that tracks "how far you went, and the calories you personally burned, if you're hiking":

> And then, there's a little record button just like you would have on your voice recording or whatever, if you're recording a video or whatever. And then you just press it to stop which is pretty neat. (P15)

P28 talked about something similar, regarding turning on and off the GPS so that:

> ...it'll mark how far you've walked. And then when you tell it to stop, it's like, 'Okay, well you've walked one mile at this pace so you burned this many calories.'

In the above instances, the user provided information to the tracker that marked a state transition from one category of activity to another, enabling her to see a representation of the data calculated by the tracker in the interface that matched her own awareness of and intention for what she was doing at that time.

In addition to specifying periods of higher activity, some trackers allow users to manually specify that they have entered sleep mode, which changes the tracker's interpretation of movements registered by the device from steps to restless sleep. P01, P05, P12, P24, and P25 all described how they manually "double tap" the tracker (or "hit it twice really fast") to make it enter sleep mode. P01 said, "You tell it when you go to sleep and you tell it when you wake up and it tracks how you were sleeping."

However, as, P05 described, this manual stopping and starting feature has some limitations:

> So for sleep you have to set it, like you have to tell it. So I'm inputting that. I'm turning it into sleep mode. I'm turning it off of sleep mode. Although I forgot to turn it off today till like one o'clock. But [laughter]... No, I did not sleep till one o'clock unfortunately, I wish.

This form of manual data entry enables the device to collect a different kind of data for that time period, like a higher activity level or calorie usage. Using an input mechanism to tell the tracker that certain data should be interpreted as being related to a particular activity helped participants to become more aware of what kinds of data the tracker can and cannot collect, and when. By thinking about data collection as something that must be started and stopped, like turning on a recording device, it supports a more limited set of expectations about what data the tracker can collect on its own.

5. DISCUSSION

Users are expected to self-manage their privacy by making choices about consent for what kinds of data collection to allow. However, they cannot do this effectively if they cannot reason about what kinds of data collection and inferences are possible. Our study focuses on folk theories, because this allows us to understand how activity tracker users think about their data, and therefore what their knowledge and experience allows them to base their privacy-related decisions on. Knowing more about their folk theories can help us better understand the boundaries between what users can and cannot reasonably consent to. Our design implications suggest ways to encourage speculation and thereby broaden users' folk theories, which could help them to better self-manage their privacy.

Our findings indicate that folk theories of activity tracker data collection arise from information provided in the interface, and from users' own perceptions of their activities. The folk theories we elicited involved three categories of data: that which users *enter* about themselves like age and weight, data that are *measured* by the tracker like steps and location, and data that are *calculated* based on other data like activity level, distance and calories burned. However, these folk theories about data types do not include other kinds of information that might be inferred from the raw sensor data generated by activity trackers, but are not directly related to activity tracking. In other words, the folk theories are constrained by what participants use activity trackers for.

Conceptualizing steps as a discrete unit of measurement, for example, supports reasoning about physical activity and fitness. But at the same time, it prevents understanding that in order to identify a step the tracker must engage in a statistical classification task. It also prevents the realization that if movement data can be used to count steps, other movements the tracker detects could be used to count other kinds of actions. This means that activity tracker users whose folk theories do not include movement as a measured data type or who do not know that steps and sleep are estimates based on movement are unlikely to be able to truly consent to the collection of data types that are calculated based on movement data. Even a belief that both steps and sleep patterns, two very different kinds of activities, are based on movement did not inspire our participants to speculate about other kinds of data that might be derived from movement.

While our participants were inspired to speculate about some aspects of the collection of certain types of activity data, there seem to be few opportunities presented by activity trackers for users to engage in the kind of speculative reasoning that generalizing beyond what the tracker was directly presenting to them would require. For example, no one who mentioned GPS, location or distance said that their tracker knows where they live, either as part of the initial free list activity or during the follow up questions and probing. This poses a problem, from a privacy perspective, for users considering whether to consent to sensor-related data collection: if users' folk theories do not include a framework for reasoning about possible inferences from sensor data, they cannot make informed choices about which systems to use and what information they do and do not want collected about them. However, our findings point to ways that interfaces might be designed to induce the kind of speculation and thinking that would engender the development of folk theories that would be more helpful for privacy-related consent decisions.

5.1 Revealing the Context of Production

Activity tracker systems involve sensor technologies, devices, apps, and cloud services that all play a part in transforming the raw sensor data into information representing actions (e.g., steps) and physiological processes (e.g., heart rate) that users can see and understand. One important input into the folk theories of participants in our study was experiences they'd had that provided visibility into how data are produced, such as seeing step counts, active minutes, and calories burned increase together in the interface, or noticing inaccuracies. However, knowing there was a heart rate sensor and seeing their heart rate in the interface did not help users in our study to become more aware of how the device is able to determine their heart rate. For example, only three participants talked about math (P15, P21) formulas (P21), or in one case, an algorithm (P23) operating on data that their trackers collected.

Raw data, or the direct output from the sensors in the activity trackers, does not have meaning by itself. It only becomes meaningful after being processed and presented to the user, in such a way that they can see themselves in their data [49]. This transformation is work that the system does on the user's behalf, so that they do not have to track their activities and perform those calculations themselves. The interface between the user and tracker hides this work, so that users are given no cause consider that step counts are not raw data. For the activity tracker users in our study, raw sensor data was not a salient aspect of *information that an activity tracker knows*. This hidden work presents a challenge for supporting folk theory development; because folk theories arise from people's experiences, users must be able to encounter or experience some aspect or evidence of this work for it to be incorporated into their folk theories.

Vertesi et al. [50] wrote about the importance of knowing the context of data production, or "how the data is crafted and acquired," in scientific collaborations. They emphasized that hiding the work that goes into preparing scientific data for sharing outside the team that produced it obscures the sociotechnical infrastructure that gives it value and meaning. In an activity tracker system, sensors, raw data, processing and other infrastructure are also invisible to the users who interact with the final output in the displays of their activity tracker devices and apps. In packaging up raw sensor data as activity data, the details of the context of production are left out in order to allow the activity data to gain credibility, resulting in processed data that seem more definitive and "true" than they really are. In other words, hiding the relationship between what the system is doing and what the user sees prevents the user from developing folk theories about data as interpretations and inferences, not absolute facts. Obscuring the ambiguity may help people become more confident in the data, but it also prevents them from speculating about what else it might be used to infer, and forming folk theories that incorporate ideas about data processing, transformation and dependencies. Information that is not incorporated into people's folk theories cannot help them to imagine potential consequences of data collection, or reason about privacy-related decisions.

5.2 Implications for Design

The seamless approach to the design of ubiquitous computing systems, as Weiser said, "focusing on the task, not the tool" [54], hides uncertainty by replacing it with certainty [9]. However, Kay and Kummerfeld [20] argue that systems should be *scrutable*, or understandable through study and observation. A scrutable system has an interface that allows the user to see the "evidence source" and the "interpretation processes" that produce the information that is consumed. Bellotti and Sellen [5], in an early paper about designing for privacy in ubiquitous computing systems, wrote about empowering users by creating designs that provide feedback about these invisible aspects. It may therefore be better for privacy to be less seamless and more scrutable; to look for ways to reveal hidden work and help users make connections between the data collection and dependencies they are already aware of in the activity tracking context, and other information that may be only indirectly related to that context.

One challenge inherent to making the production of activity tracker data more observable is that users may find the additional information overwhelming and not know what to do with it. For example, Rapp and Cena [39] found that people who had never used activity trackers before participating in their study felt the data and graphs the trackers provided were already "too abstract and removed from what they were expecting", not meaningful to them, and difficult to engage with. However, our findings suggest several ways that small design changes to the information provided in the tracker's interface might support the development of folk theories through encouraging speculation about how the data are produced.

Seeing simultaneous changes to multiple data types in the app interface (e.g., steps and calories burned) led to folk theories that incorporated causal relationships between those data types. But, participants needed a reason to be looking at the interface in the first place in order to see the relationship between those data types, and that reason is activity tracking. Presenting information about other kinds of data dependencies that are related to but not directly about activity tracking may be a minor departure from the user's main task that engenders speculation about what else an activity tracker might know.

Many services based on sensor data periodically publish essays on the company's blog or website providing analysis of patterns in the data generated by users of the service; Fitbit is one example of this[4]. If activity tracker service providers were to incorporate information comparing users' data with aggregate statistics as part of the app's interface, it could provide additional visibility into the aggregation that underlies all of the data output that users interact with. For example, when reporting sleep patterns, the app could also present information like, "Your average bed time is 11:23 PM, which is 20 minutes later than other users in your age group." Alternately, to promote awareness of the possibility that a user's location might be used to generate new data about semantic aspects of geography such as where the user lives, the tracker could display to the user information about how far the participant went from home that day while jogging (not just length of the run), or how far from home their number of steps that day would have taken them. In a more "creepy" vein [44], an activity tracker app might inform users that "people who have restless nights that are similar to yours are likely to be new parents." Folk theories incorporating the kinds of insights that can be derived through aggregation might allow users to consider consequences like this when reasoning about possible privacy-related effects of using sensor-enabled technologies.

Tracking an activity that users can't directly perceive, like sleep, led to doubt and speculation from our participants about how the tracker could measure a phenomena like this. Sleep is unique in the context of activity tracking, in that it is the only activity that is not verifiable by the user while it is happening. However, other kinds of activities that might be detected also have this characteristic, to varying degrees. For example, Fitbit trackers began providing information about "stationary time", or amount of time spent without moving in a given time interval, to users in April 2016, after our study was conducted. It may be difficult for users to pay attention to the absence of an activity, but trackers can do this easily. It therefore might be possible to combine information about stationary time with GPS, and highlight data types in the interface like time spent sitting at work, or in a moving vehicle. Making these data visible could encourage users to think about how the tracker defines "stationary time," how the data are collected, how the location categories are defined, and how different data types can be combined to produce new data.

In tracker devices with an optical heart rate sensor, *visibility of the sensor component* made the source of the data collected by the device more salient for our participants, and changed the way they reasoned about the data. With the current trend towards making trackers look less like fitness equipment and more like clothing accessories, making additional sensors more visible seems like an unlikely possibility. However, perhaps there is a way to make the raw data more conceptually tangible. It might be possible to quantify aspects of the tracker itself, like the tracker quantifies aspects of the person. Many personal computers include widgets and control panels that present statistics about the "health" of the device, such as available memory, temperature, fan speed, and uptime. Similar kinds of data could be calculated about the tracker device, or about user interactions with the tracker. For example, data about how many times the user has checked the tracker's display in the last week might make the device more salient to the user in ways that are both informative and provide a focus on technical details for users to speculate about and incorporate into their folk theories.

As others have found [13, 56], *perceived inaccuracy* prompts attention to aspects of how the data are collected. It may also present a view into the statistical model that data like step counts are based on. This was a powerful mechanism supporting speculation among the users in our study about how their trackers counted steps and measured sleep. However, this speculation only extended to data types they knew the tracker was supposed to be collecting. Perceived inaccuracy highlights uncertainty in the underlying machine

[4] `https://blog.fitbit.com/how-do-your-sleep-habits-stack-up/` (accessed on July 10, 2017)

learning models, and therefore is a direct way to encourage users to notice and think about the context of data production. Folk theories that incorporate concepts related to the production of data may help users to reason about inferences and calculated data types.

Consolvo et al. [10] wrote that it is important for future research to consider better ways to present uncertainty to the user, and to understand its effects on user behavior. However, they also said that this is challenging, because typical ways of presenting statistical uncertainty are unlikely to be understood by most users. The challenge for design to support folk theories of data collection is how to provide information that helps the user connect the realization that a tracker may be collecting some kind of data other than steps, to specific other kinds of information or activities the tracker might be able to detect. "Glanceable" displays on activity trackers with a smart watch form factor traditionally have been focused on presenting status updates related to activity over the past hour, goal attainment, etc. [15]. But it might be possible to use the displays to notify users about some of the uncertainty involved in activity recognition, by using colors or shapes to indicate deviation from the underlying statistical models.

Finally, *manual "recording" of activities* via state transitions that are entered by the user, like activating sleep mode, are also a form of data collection. These data give the tracker additional context to use to interpret the raw sensor data collected during certain time periods. The act of starting and stopping the "recording" also gives the user more confidence in the accuracy of the data that are collected about the activity. If the user were able to provide other kinds of contextual information to the tracker, it could both help improve the functioning of the system, and also help the users better understand the context of production. For example, activity tracker users could be given the ability to contribute data consisting of feedback on the tracker's performance. A "thumbs up" or "thumbs down" might signal points at which they feel the tracker is particularly accurate or inaccurate. Data like this collected over time and relayed back to the user in aggregate might provide visibility into the messiness of the context of production and the work that goes into estimating step counts, while also providing information that users would find helpful for understanding when they can trust the tracker and when they cannot, and that system operators would find helpful for improving accuracy.

5.3 Implications for Privacy Self-Management

Design to encourage speculation about the context of production of activity tracker data has implications for the formation of folk theories about sensor data collection, and for helping users make decisions about privacy self-management and consent. Folk theories are "ways of understanding" [12] that are based on experience and help users of technologies make decisions [53]. In other words, folk theories are cognitive structures that help users envision what might happen based on what they already know. A folk theory that includes knowledge about sensors and the kinds of data an activity tracker records about the world, or the concept that the numbers displayed by the tracker are estimates with a degree of uncertainty, or that some data are produced by combining other types of data, may help users to speculate and imagine different possible consequences than a folk theory that involves certainty that steps are directly counted.

This does not mean that folk theories need to be technically accurate from an expert's perspective. Kempton demonstrated in his thermostat study that incorrect mental models about technology can still be useful for decision-making; in his case, for making home heating decisions [22]. It is not necessary for a user to understand how an accelerometer works, or what the algorithm for identifying a step is, to speculate that if step counts are estimates other kinds of information may be estimated too. Speculation does not need to produce accurate knowledge to be useful for reasoning about control over data collection and possible consequences. Folk theories that do not help users reason about possible consequences beyond health and fitness may not be helpful for making consent choices about data collection in systems that involve inferences beyond the direct context of use. Folk theories that involve speculation about aspects of the context of production could provide better support for informed privacy self-management and consent.

6. CONCLUSION

Sensor-enabled systems, like activity trackers, collect highly detailed and personal data about users' behavior. Because people are expected to be able to self-manage their privacy regarding digital information, it is important to understand users' folk theories of this sensor data collection, which help them reason about new situations and make decisions. Our findings show that users' folk theories are limited to the activity tracking context, and do not help users reason about other kinds of data that might be collected or used beyond activity tracking. Instead, activity tracker interfaces obscure the complexity and uncertainty involved with producing the data that are shown to users. By hiding the messiness of transforming raw data into useful insights, the data that are collected become more helpful for the user's primary task (health and fitness), but not useful for reasoning about privacy, which is at best a background task.

Despite this, users have experiences with their trackers that open them up to speculating about how their data are produced, and to learning about connections between data types. While designs that provide hints about some of the complexity may come with some cost for the user, our findings suggest avenues for design that build on speculation users are already engaged in, in ways that are peripherally related to current tracker functionality. Future work is needed to further understand the connection between speculation, folk theories about data collection, and user reasoning about privacy and consent.

7. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. CNS-1524296.

8. REFERENCES

[1] H. Almuhimedi, F. Schaub, N. Sadeh, I. Adjerid, A. Acquisti, J. Gluck, L. F. Cranor, and Y. Agarwal. Your Location has been Shared 5,398 Times! In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 787–796, 2015.

[2] F. Asgharpour, D. Liu, and L. J. Camp. Mental Models of Security Risks. In *International Conference on Financial Cryptography and Data Security*, pages 367–377, 2007.

[3] R. Balebako, J. Jung, W. Lu, L. F. Cranor, and C. Nguyen. "Little brothers watching you:" raising awareness of data leaks on smartphones. In *Symposium on Usable Privacy and Security*, page Article 12, 2013.

[4] P. Bayman and R. E. Mayer. Instructional Manipulation of Users Mental Models for Electronic Calculators. *International Journal of Man-Machine Studies*, 20(2):189–199, 1984.

[5] V. Bellotti and A. Sellen. Design for Privacy in Ubiquitous Computing Environments. In *Proceedings of the European*

Conference on Computer-Supported Cooperative Work, pages 77–92, 1993.

[6] H. R. Bernard and G. Ryan. *Analyzing Qualitative Data: Systematic Approaches*. SAGE Publications, 1st edition, 2009.

[7] S. P. Borgatti. Elicitation techniques for cultural domain analysis. In J. Schensul and M. LeCompte, editors, *The Ethnographer's Toolkit, Vol. 3*, pages 115–151. Altimira Press, 1998.

[8] D. D. Brewer. Supplementary Interviewing Techniques to Maximize Output in Free Listing Tasks. *Field Methods*, 14(1):108–118, 2002.

[9] M. Chalmers, I. MacColl, and M. Bell. Seamful design: Showing the seams in wearable computing. In *Eurowearable*, pages 11–16, 2003.

[10] S. Consolvo, P. Klasnja, D. W. McDonald, and J. A. Landay. Designing for Healthy Lifestyles: Design Considerations for Mobile Technologies to Encourage Consumer Health and Wellness. *Foundations and Trends in Human-Computer Interaction*, 6(3-4):167–315, 2014.

[11] M. Eslami, K. Karahalios, C. Sandvig, K. Vaccaro, A. Rickman, K. Hamilton, and A. Krilik. First I Like It, Then I Hide It: Folk Theories of Social Feeds. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2371–2382, 2016.

[12] S. A. Gelman and C. H. Legare. Concepts and Folk Theories. *Annual Review of Anthropology*, 40:379–398, 2011.

[13] N. Gorm and I. Shklovski. Steps, Choices and Moral Accounting: Observations from a Step-Counting Campaign in the Workplace. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 148–159, 2016.

[14] S. D. Gosling, S. Vazire, S. Srivastava, and O. P. John. Should We Trust Web-Based Studies? A Comparative Analysis of Six Preconceptions About Internet Questionnaires. *American Psychologist*, 59(2):93–104, 2004.

[15] R. Gouveia, F. Pereira, E. Karapanos, S. A. Munson, and M. Hassenzahl. Exploring the design space of glanceable feedback for physical activity trackers. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 144–155, 2016.

[16] B. Hulkower. Fitness Fuels US Consumer Desire for Wearable Tech, 2016. `http://www.mintel.com/press-centre/technology-press-centre/fitness-fuels-us-consumer-desire-for-wearable-tech-with-sales-increasing-186-in-2015`.

[17] A. Jackson. Husband and Wife Never Expected Their Fitbit Would Tell Them This, Feb. 2016. `http://www.cnn.com/2016/02/10/health/fitbit-reddit-pregnancy-irpt/`.

[18] P. N. Johnson-Laird. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250, 2010.

[19] R. Kang, L. Dabbish, N. Fruchter, and S. 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*, pages 39–52, 2015.

[20] J. Kay and B. Kummerfeld. Creating personalized systems that people can scrutinize and control: Drivers, principles and experience. *ACM Transactions on Interactive Intelligent Systems*, 2(4):Article 24, 2012.

[21] F. C. Keil. The Feasibility of Folk Science. *Cognitive Science*, 34(5):826–862, 2010.

[22] W. Kempton. Two Theories of Home Heat Control. *Cognitive Science: A Multidisciplinary Journal*, 10(1):75–90, 1986.

[23] P. Klasnja, S. Consolvo, T. Choudhury, R. Beckwith, and J. Hightower. Exploring Privacy Concerns about Personal Sensing. In *International Conference on Pervasive Computing*, pages 176–183, Mar. 2009.

[24] J. Lin, S. Amini, J. 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 ACM Conference on Ubiquitous Computing*, pages 501–510, 2012.

[25] T. R. Lindlof and B. C. Taylor. *Qualitative Communication Research Methods*. Sage Publications, 2nd edition, 2002.

[26] T. Matlock, S. C. Castro, M. Fleming, T. M. Gann, and P. P. Maglio. Spatial Metaphors of Web Use. *Spatial Cognition and Computation*, 14(4):306–320, 2014.

[27] J. A. Maxwell. *Qualitative Research Design: An Interactive Approach*, volume 41 of *Applied Social Research Methods*. Sage Publications, 3rd edition, 2012.

[28] M. B. Miles and M. A. Huberman. *Qualitative Data Analysis: An Expanded Sourcebook*. Sage Publications, 2nd edition, 1994.

[29] V. G. Motti and K. Caine. Users' Privacy Concerns About Wearables. In *International Conference on Financial Cryptography and Data Security*, pages 231–244. 2015.

[30] H. Nissenbaum. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2009.

[31] G. Novick. Is there a bias against telephone interviews in qualitative research? *Research in Nursing & Health*, 31(4):391–398, 2008.

[32] G. Paolacci, J. Chandler, and L. N. Stern. Running Experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, 5(5):411–419, 2010.

[33] S. Petronio. *Boundaries of Privacy: Dialectics of Disclosure*. State University of New York Press, 2002.

[34] E. S. Poole, M. Chetty, R. E. Grinter, and W. K. Edwards. More than meets the eye: Transforming the User Experience of Home Network Management. In *ACM Conference on Designing Interactive Systems*, pages 455–464, 2008.

[35] E. S. Poole, C. A. Le Dantec, J. R. Eagan, and W. K. Edwards. Reflecting on the Invisible: Understanding End-User Perceptions of Ubiquitous Computing. In *Proceedings of the International Conference on Ubiquitous Computing*, pages 192–201, 2008.

[36] M. Quinlan. Considerations for Collecting Freelists in the Field: Examples from Ethobotany. *Field Methods*, 17(3):219–234, 2005.

[37] E. Rader and R. Gray. Understanding User Beliefs About Algorithmic Curation in the Facebook News Feed. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 173–182, 2015.

[38] A. Raij, A. Ghosh, S. Kumar, and M. Srivastava. Privacy Risks Emerging from the Adoption of Innocuous Wearable Sensors in the Mobile Environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 11–20, 2011.

[39] A. Rapp and F. Cena. Personal Informatics for Everyday Life: How Users Without Prior Self-Tracking Experience

Engage with Personal Data. *International Journal of Human-Computer Studies*, 94:1–17, 2016.

[40] A. K. Romney, D. D. Brewer, and W. H. Batchelder. Prediciting Clustering From Semantic Structure. *Psychological Science*, 4(1):28–34, 1993.

[41] T. A. Schwandt. *Qualitative Inquiry: A Dictionary of Terms*. Sage Publications, 1997.

[42] F. Shih, I. Liccardi, and D. Weitzner. Privacy Tipping Points in Smartphones Privacy Preferences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 807–816, 2015.

[43] P. C. Shih, K. Han, E. S. Poole, M. B. Rosson, and J. M. Carroll. Use and Adoption Challenges of Wearable Activity Trackers. In *iConference Proceedings*, 2015.

[44] I. Shklovski, S. D. Mainwaring, H. H. Skúladóttir, and H. 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*, pages 2347–2356, 2014.

[45] D. J. Solove. Privacy self-management and the consent dilemma. *Harvard Law Review*, 126(7):1880–1903, 2013.

[46] J. E. Sturges and K. J. Hanrahan. Comparing telephone and face-to-face qualitative interviewing: a research note. *Qualitative Research*, 4(1):107–118, 2004.

[47] The White House. Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy. *Journal of Privacy and Confidentiality*, 4(2):Article 5, 2012.

[48] E. C. Thompson and Z. Juan. Comparative Cultural Salience: Measures Using Free-List Data. *Field Methods*, 18(4):398–412, 2006.

[49] P. Tolmie, A. Crabtree, T. Rodden, J. A. Colley, and E. A. Luger. "This has to be the cats:" Personal Data Legibility in Networked Sensing Systems. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 490–501, 2016.

[50] J. Vertesi and P. Dourish. The Value of Data: Considering the Context of Production in Data Economies. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 533–542, 2011.

[51] R. Want, A. Hopper, V. Falcão, and J. Gibbons. The Active Badge Location System. *ACM Trans. Inf. Syst.*, 10(1):91–102, 1992.

[52] W. H. Ware and the Secretary's Advisory Committee on Automated Personal Data Systems. *Records, Computers, and the Rights of Citizens*. US Department of Health, Education & Welfare, 1973. `https://www.justice.gov/opcl/docs/rec-com-rights.pdf`.

[53] R. Wash. Folk Models of Home Computer Security. In *Symposium on Usable Privacy and Security*, page Article 11, 2010.

[54] M. Weiser. The world is not a desktop. *interactions*, 1(1):7–8, 1994.

[55] S. C. Weller and A. K. Romney. *Systematic Data Collection*. Qualitative Research Methods Series 10. Sage Publications, 1st edition, 1988.

[56] R. Yang, E. Shin, M. W. Newman, and M. S. Ackerman. When Fitness Trackers Don't 'Fit': End-User Difficulties in the Assessment of Personal Tracking Device Accuracy. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 623–634, 2015.

[57] Y. Yao, D. Lo Re, and Y. Wang. Folk Models of Online Behavioral Advertising. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1957–1969, 2017.

Replication: Challenges in Using Data Logs to Validate Phishing Detection Ability Metrics

Casey Inez Canfield*, Alex Davis*, Baruch Fischhoff*,
Alain Forget**, Sarah Pearman*, Jeremy Thomas*
*Carnegie Mellon University, **Google
caseycan@gmail.com, alexdavis@cmu.edu, baruch@cmu.edu,
aforget@google.com, spearman@cmu.edu, thomasjm@cmu.edu

ABSTRACT

The Security Behavior Observatory (SBO) is a longitudinal field-study of computer security habits that provides a novel dataset for validating computer security metrics. This paper demonstrates a new strategy for validating phishing detection ability metrics by comparing performance on a phishing signal detection task with data logs found in the SBO. We report: (1) a test of the robustness of performance on the signal detection task by replicating Canfield, Fischhoff, and Davis (2016), (2) an assessment of the task's construct validity, and (3) evaluation of its predictive validity using data logs. We find that members of the SBO sample had similar signal detection ability compared to members of the previous mTurk sample and that performance on the task correlated with the Security Behavior Intentions Scale (SeBIS). However, there was no evidence of predictive validity, as the signal detection task performance was unrelated to computer security outcomes in the SBO, including the presence of malicious software, URLs, and files. We discuss the implications of these findings and the challenges of comparing behavior on structured experimental tasks to behavior in complex real-world settings.

1. INTRODUCTION

Maintaining security on a home computer requires knowing which security practices are most important [18] and implementing those practices, even when they may be inconsistent with users' mental models of computer security [3, 43, 44]. Users are expected to keep their software up to date (both individual programs and their operating system), avoid suspicious links and attachments (i.e. phishing attacks), choose secure passwords, and install security programs (e.g. anti-virus). Many struggle to understand and follow all these recommendations, despite good intentions.

Meanwhile, cyberattacks are becoming more varied and pervasive [39, 40], where about 1 in every 2,600 emails are phishing attacks (primarily targeted spear phishing attacks), resulting in losses of over $3 billion from business email compromise scams over the last three years [39]. Phishing attacks are no longer limited to email, but can occur over instant messenger, social media, or text messages [39]. Phishing is often used to introduce malware to a computer [37], resulting in prolonged risk. Although there are products to help protect users, none are perfect. For example, email providers use spam filters, browsers employ blacklists to block malicious websites, and security programs block and delete malicious files and software. In some cases, this requires user engagement, such as updating security programs (if automatic updating is not enabled). In other cases, such as browser blacklists, users have little control.

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12 -- 14, 2017, Santa Clara, California.

Growing concern over phishing risks is driving the need for timely, cost-effective measures of individuals' vulnerability. Such metrics might be derived from actual behavior or a dedicated test. Any metric faces three challenges: (a) it must differentiate between users' ability (e.g. to detect phishing emails and maintain software) and the technology in place to protect them (e.g. spam filters and blacklists, automatic updates); (b) it must account for the low base rate of phishing attacks; and (c) it must be able to extrapolate from the observed circumstances to those where users are faced with actual attacks. A simple test with predictive validity could guide targeted interventions if it provided useful performance measures.

Here, we demonstrate a new strategy for validating metrics, by triangulating performance on an experimental task with real-world system outcomes. The experimental task was developed by Canfield, Fischhoff, and Davis [4] (referred to as Canfield et al.). It extracts individual-level signal detection measures of phishing vulnerability and was demonstrated with an online mTurk sample [4]. We validate these measures using the *Security Behavior Observatory (SBO)*, a longitudinal field study that provides detailed data on a community sample of computer users' security habits over time [9, 10].

Signal detection theory (SDT), when applied to phishing detection, distinguishes between users' ability to tell the difference between phishing and legitimate emails (sensitivity or d') and bias toward identifying ambiguous emails as phishing or legitimate (response bias or c) [24]. SDT is more useful than other metrics, such as accuracy, because it accounts for the tradeoffs that people make between false negatives (missing phishing emails and potentially falling for an attack) and false positives (mistaking legitimate emails for phishing by deleting an important message or reducing the efficiency of email).

Here, we first replicate Experiment 1 from Canfield et al. with SBO participants in order to assess the robustness of their original (mTurk) results [4]. We then assess the construct validity of those performance metrics, in terms of correlations with self-reports on the Security Behavior Intentions Scale (SeBIS) [6]. Finally, we assess their predictive validity in comparisons with evidence of security vulnerabilities on their home computers.

Canfield et al.'s task uses realistic email messages to elicit users' detection ability and behavior. Their experiments found robust results across several experimental manipulations in mTurk samples [4]. As described below, the SBO sample is different in many ways, raising the question of how similar their performance will be. If it proves robust, one can then ask how strongly it is related to other computer security performance measures. Experimental measures are extracted under conditions where participants know that they are being observed, which may affect their behavior in various ways, including behaving so as to satisfy (or perhaps frustrate) perceived research goals [31, 35]. Here, we consider two such tests:

- *Construct validity* [5]: how well the SDT measures correlate with another theoretically related (and validated) measure, the Security Behavior Intentions Scale (SeBIS) [6]; and
- *Predictive validity*: how well the SDT measures predict actual behavior, tested by whether they improve the fit of logistic models for vulnerability to phishing attacks.

We find some evidence of robustness and construct validity, but not for predictive validity. We discuss the ways in which that failure reflects on the measures and on the challenges of characterizing vulnerability in real-world settings, shaped by users' behavior (regarding security and other matters) and their computer environment (e.g. browser, OS).

2. BACKGROUND AND RELATED WORK

The probability of users experiencing negative computer security outcomes (such as viruses) reflects both their vulnerability and their exposure [26]. The former includes their ability to detect and avoid threats (e.g. identify phishing emails), as well as their engagement in risky behavior (e.g. not updating software). The latter reflects their attractiveness as targets. We now review research on (a) measuring phishing detection performance and (b) determinants of vulnerability and exposure.

2.1 Measuring Phishing Detection Performance

There are two primary strategies for measuring users' phishing detection performance: subjective and objective. Egelman and Peer developed a subjective scale of Security Behavior Intentions (SeBIS), with four subscales: device securement, password generation, proactive awareness, and updating [6]. The proactive awareness subscale, which measures attention to URLs, has special interest for phishing vulnerability. Low scores on the proactive awareness subscale have been related to impulsivity, risk-taking, and dependence (i.e. relying on other people), consistent with the phishing detection literature [6, 34, 42, 47]. In a test of validity, Egelman, Harbach & Peer found that performance on the proactive awareness scale was correlated with respondents' ability to detect a phishing website in a laboratory environment without priming (without telling them that they were being tested on that ability) [7]. The only way to determine whether it was a phishing website was to look at the URL. Although only 22 of 718 participants correctly identified the phishing website, their proactive awareness scores were significantly higher than those of the rest of the sample [7].

Objective measures assess users' actual ability to identify phishing emails, rather than relying on self-reports of how well they do. They allow varying experimental conditions, to examine the effects of situational factors (e.g. perceived consequences, habits, stress) on phishing vulnerability [4, 41]. Canfield et al. estimated SDT parameters on an individual level by asking users to identify which emails were phishing in a set of 38 in an online test [4]. However, such tests are vulnerable to experimenter demand effects, where subjects do what they think they should do, rather than what they would normally do [27, 38].

Although controlled studies are cheaper and easier to implement than field tests, it is important to validate such measures against real world behavior. Ideally, one would send emails (both legitimate and phishing) to participants to determine how well their performance in the artificial test environment reflects their normal behavior. However, this is not always possible. It can be challenging to do high-fidelity field tests (i.e. without providing feedback on performance) without putting users at risk or incurring high costs. (Examples include Jakobsson et al. [19] and Kumaraguru et al. [20]).

As an alternative to field tests, Sotirakopoulos et al. propose examining logs of user behavior [38]. The SBO is an ongoing data collection effort that collects such logs. Its wealth of data provides multiple ways to assess vulnerability and to account for other factors that might influence users' experiences of negative computer security outcomes. Demand effects are expected to be minimal given that the study software is sufficiently unobtrusive that participants often report having forgotten that they were participating in the study. In a similar observational study, few participants reported altering their behavior in an exit survey [21].

2.2 Determinants of Vulnerability and Exposure

Unsophisticated or careless users may escape harm if they seldom use their computers or avoid dangerous situations. Conversely, knowledgeable users may ward off most attacks, yet still succumb if they use their computers heavily or are valuable targets, subject to particularly effective attacks (such as spear phishing).

Research suggests that user knowledge alone cannot compensate for the increased odds of exposure to negative computer security outcomes that come with increased use. Although research on phishing susceptibility has found that individuals with higher computer literacy are less susceptible to individual phishing attacks [36, 47], more computer-literate users tend to use their computers more frequently [2], increasing their chances of exposure to attacks and negative outcomes [22]. SBO research suggests that users' engagement with security issues, as expressed in interviews, is not a good predictor of their security outcomes [10]. Lalonde-Levesque et al. also found that more technically-savvy users are more likely to be exposed to malware threats [21]. Therefore, it is critical to control for exposure when assessing the relationship between phishing detection performance and negative computer security outcomes.

Users may also experience more negative computer security outcomes because they engage in risky behavior, such as frequently clicking on links in emails or not updating their anti-virus software. In a survey of Dutch citizens, Leukfeldt found that while the OS type was related to malware, updated anti-virus was not [22]. While anti-virus software protection against social engineering and zero-day exploits is limited, one would expect protection against spam-type attacks using known malicious software. Our analysis assesses the relevance of this variable.

3. METHOD

3.1 Decisions in Phishing Scenarios (SDT)

Canfield et al. [4] used a scenario-based approach [20, 34], in which participants reviewed emails of a fictitious persona, Kelly Harmon. Before beginning that task, participants reviewed the PhishGuru comic strip [29] to ensure that they had some knowledge of phishing and understood their task. They then saw one of two *notifications of base rate*: "Approximately half of the emails are phishing emails" or "Phishing emails are included." *Attention* was a binary (0,1) measure, where 1 described participants who correctly answered 3 questions: "Where does Kelly Harmon work?", "What is a phishing email?", and an email that said, "If you are reading this, please answer that this is a phishing email."

Participants evaluated 38 email messages, half of which were phishing (adapted from public archives), in a random order. The base rate of phishing emails (50%) was much higher than in everyday settings (<1%) [39] in order to collect enough judgments without overburdening participants. We used the same stimuli as Canfield et al. [4] (available online at https://osf.io/7bx3n/). They ranged in difficulty from obvious phishing messages with typos to more sophisticated spear phishing attacks. For each email, participants answered the following questions:

1. *detection*: "Is this a phishing email?" (Yes/No);
2. *behavior*: "What would you do if you received this email?", with multiple-choice options including "click link/open attachment," "check sender," "check link," "reply," "ignore or archive it," "delete it," "report as spam," and "other" (following [36]);
3. *confidence*: "How confident are you in your answer?" (50-100%); and
4. *perceived consequences*: "If this was a phishing email and you fell for it, how bad would the consequences be?" (Likert scale: 1 = not bad at all to 5 = very bad).

We limited the replication to Experiment 1 in Canfield et al., which asked all participants to perform both the detection and behavior tasks. In Experiment 2, participants were randomly assigned to perform either the detection or behavior task. Canfield et al. found no significant differences in the SDT performance metrics between Experiment 1 and Experiment 2. Given the limited sample of SBO participants, having all participants perform both the detection and the behavior tasks maximized the precision of our parameter estimates. We also measured the time spent on the phishing information comic and median time spent on each email. Finally, we collected demographic information on gender, age, and education.

We evaluated individual performance using signal detection theory (SDT), a mathematical method for characterizing users' ability to distinguish phishing and legitimate emails (d') and their bias toward perceiving emails as phishing or legitimate (c). The SDT measures capture the trade-off between hit rates (*H*, correctly identifying emails as phishing) and false-alarm rates (*FA*, incorrectly identifying legitimate emails as phishing) using an inverse normal transformation to convert the probability to a Z-score:

$$d' = z(H) - z(FA)$$
$$c = -0.5(z(H) + z(FA))$$

As described by Canfield et al. [4], we estimated SDT parameters for the detection ($_D$, question (1) above) and behavior ($_B$, question (2) above) tasks separately. Thus, we calculated four phishing vulnerability parameters, summarized in Table 1.

Table 1. Phishing vulnerability parameters calculated using signal detection theory (SDT) for replication and validation of Canfield et al. [4].

Parameter	Definition
Detection Sensitivity (d'_D)	Ability to distinguish between phishing and legitimate emails.
Behavior Sensitivity (d'_B)	Ability to distinguish between when to click on links and when not to.
Detection Response Bias (c_D)	Bias toward identifying an email as phishing (negative c) or legitimate (positive c).
Behavior Response Bias (c_B)	Bias toward clicking on links (positive c) or not (negative c).

3.2 Security Behavior Intentions Scale (SeBIS)

As part of their SBO tasks, 84 participants completed the Security Behavior Intentions Scale (SeBIS) [6]. The SeBIS has 16 statements describing behaviors divided into four subscales: device securement, password generation, proactive awareness, and updating. Respondents rate on a Likert scale whether they *never* (1) to *always* (5) perform the stated behavior. Conceptually, the signal detection measures should be most closely related to the proactive awareness subscale, which includes five statements related to evaluating links, such as "When browsing websites, I mouseover links to see where they go, before clicking on them" and "I know what website I'm visiting based on its look and feel, rather than by looking at the URL bar" (reverse coded).

3.3 Home Computer Security Outcomes (SBO)

The Security Behavior Observatory (SBO) is an ongoing longitudinal study, gathering field data about home users' computer security habits. SBO participants agree to install the project software on their personal computers to gather data on their Internet browsing, installed applications, processes, network connections, system events, and more. This software then securely transmits the data to the researchers.

From these data, we measured three types of negative computer security outcomes: (a) visits to malicious URLs, (b) installed malware, and (c) presence of malicious files. Malicious URLs were identified using the Google Safe Browsing API [14] with participants' web browsing (i.e. Internet Explorer, Chrome, and Firefox) and network packet data. Due to technical limitations with browser extensions, we were unable to collect data from other popular browsers, such as Microsoft Edge. However, those data were observed in the network packet data, which include all HTTP traffic for each webpage, making it a much richer source than the browser data, which only record webpage URLs. The average webpage has approximately 100 HTTP requests for the

HTML, CSS, images, ads, multimedia, JavaScript, Flash and other files that form a single webpage [17].

We identified malware with ShouldIRemoveIt.com, which is designed to help users remove unwanted applications from their computer. We identified malicious files with VirusTotal.com, a subsidiary of Google that aggregates anti-virus scanners. For flagging software or files as malicious, we used a threshold of at least 2 scanners for ShouldIRemoveIt.com and at least 2% of scanners for VirusTotal.com. Using greater scanner agreement did not significantly change the results. Malicious files were identified across the entire computer, while malware was limited to installed applications. We assessed each outcome as a binary variable (where 1 indicates that the outcome was observed at least once and 0 indicates that the outcome was not observed), rather than a continuous one (i.e. number of negative outcomes) due to the high number of participants who had no negative outcomes (i.e. had never visited a malicious website or had no malware) and the potential unreliability of count data [23]. Participants varied in how much they used their computers, which as described above is related to the observation of negative outcomes.

We constructed logistic regression models for each outcome following the logistic model construction strategy outlined by Hosmer et al. [16] for identifying potential predictors, defined as those with statistically significant univariate correlations with the outcomes. These potential predictors are described in the next two subsections. To avoid bias and maintain transparency, we preregistered the logistic regression models at the Open Science Framework (https://osf.io/jkhbv/) before combining the SBO and SDT experiment data [25, 28]. The analysis reported here differs from the proposed analysis in the preregistration due to our acquiring more SBO data. We also improved the analysis by: (a) eliminating repetitive measures (e.g. counts of social media domains), (b) implementing an automated process for identifying malware, rather than relying on manually coded items, and (c) adding malicious files as an outcome variable.

3.3.1 Browsing exposure and risky behavior

We identified 3 variables to describe browsing exposure. Each was calculated separately for the browser and network packet data. They were (a) *total URLs/day*, (b) *unique URLs/day*, and (c) *domains/day*. Each daily count was only for days that data were received from the participant's machine.

We measured risky behavior in terms of counts of *clicked email links/day*. We expected users who clicked on more links in emails to be more likely to visit malicious URLs. We assessed this activity in 2 ways: (a) URL tracking, for URLs that include "mail" or "email" after =, &, or ? (excluding email domains), and (b) source data, where the source URL was an email domain and the destination was not. The source data did not describe whether links were clicked from an email software client, such as Microsoft Office Outlook. For the network packet data, we could only use the URL tracking method (a), because source data were unavailable.

3.3.2 Software exposure and risky behavior

We measured software exposure as a count of *total software*, excluding updates, installers, and language packages.

We sought to measure risky behavior with three variables: *delayed software updates*, *days since Windows update*, and *third-party security software* (e.g. anti-virus, anti-malware). Delayed software updates on SBO participants' computers is a count, ranging from 0 to 6, of the number of outdated popular software including Adobe Flash, Adobe Reader, Java, Internet Explorer, Chrome, and Firefox. A program was considered outdated if the participant's computer had not updated to the latest version a week after it was released. Days since Windows update is the number of days since a Windows update was most recently installed. Thus, a low number suggests that the user has updated their Windows OS more recently. This measure does not capture why users waited to install updates (e.g. whether they actively delayed updates or did not see prompts).

For third-party security software, we assigned a binary variable where 1 indicated that it was installed and error-free (see below) and 0 indicated errors or no software. Security software was considered error-free if it was in use for over 7 days, updating without errors, and scanning. In some cases, it was impossible to know if a security program met all these criteria because either it did not log the data or the log was not informative. In those cases, we used the available subset of these criteria. Thus, we assumed that installed security software was error-free unless there was evidence otherwise. We could examine the logs for McAfee, Malwarebytes, Webroot, Avast, Norton, Kaspersky, and AVG to assess their median days in use: 172 (M = 223, SD = 238). We could not assess updating for Avast or scanning for McAfee, Avast, and AVG due to missing or uninformative logs.

3.4 Sample

SBO participants were recruited from local participant pools and are predominantly retirees and college students. For this study, we recruited participants from among those who joined the SBO between October 2015 and February 2016, asking for volunteers to participate in "an online research study about email use." In addition to their regular monthly SBO compensation, each received $20 upon completing our phishing detection experiment. Those who did not start the experiment were sent a reminder after 9 days. Those who started, but did not finish, were sent a reminder after 9 days and again after another 7 days. SBO participants received higher compensation than mTurk participants ($20 vs. $5) to encourage a high response rate, given the limited pool of SBO participants. This study was approved by the Carnegie Mellon University Internal Review Board.

3.5 Defining Successful Replication

The replicability of Canfield et al. can be assessed in terms of the methods (also referred to as reproducibility) and results [13]. Canfield et al. made their original study materials and code publicly available[1] and this paper follows suit to ensure the methods are reproducible (see Appendix). The following analysis is focused on assessing whether the results are replicable and robust to changes in the study sample.

There is an ongoing debate regarding how to measure a successful replication [1, 8, 11, 29]. For this study, we assess whether the replication was successful in four ways:

1. Comparison of effect sizes
2. Consistency of the hypothesis test results
3. Parameter space region ruled out by confidence intervals
4. Combined analysis

We (1) directly compare the point estimates or effect sizes of the SDT parameters for the original and replication study. First we qualitatively compare the point estimates, considering a meaningful difference of a 10% change in the hit rate, or probability of detecting phishing emails as unsuccessful replication. For the SDT parameters, this is a difference of 0.3 for d' and 0.1 for c. We then use a two-sample statistical significance test of the null hypothesis that the two studies were drawn from populations with the same effect size. The limitation of this first approach is that a conclusion that the study replicated based on the failure to reject the null hypothesis depends on the statistical power of the test, and thus sample size of both studies. Lower statistical power would lead to a higher frequency of conclusions that the study replicated even in the face of large differences, and high statistical power would lead to conclusions that the study did not replicate even if the differences in effect sizes were trivial.

Our second test (2) assesses the consistency of the regression coefficients in the replication study with the null hypothesis that the regression coefficient is exactly zero. The p-value on the t-test of each regression coefficient provides this measure of consistency [45]. If the p-value is below the .05 alpha level, we conclude that the regression coefficient from the replication study is inconsistent with zero, and that the study successfully replicated. The limitation of this second approach is the opposite of the first, where lower statistical power would lead to fewer conclusions that the study successfully replicated even if the regression coefficient was large, and high statistical power would lead to more conclusions that the study successfully replicated even if the regression coefficient was small.

Third, we assess (3) the region of the parameter space ruled out by confidence intervals. In the original and replication studies, we construct 95% confidence intervals. Each interval either does or does not cover the population parameter, and if we conclude that it does include the population parameter, then we will be wrong 5% of the time (i.e. the population parameter falls outside the interval). Therefore, a successful replication would find similar conclusions about the population parameter (i.e. that the region of the parameter space outside the interval in the two studies is "similar"). We operationalize this similarity as having a non-empty union of the two intervals, or that the intervals overlap. In other words, we judge that a study replicated the first if the two studies do not rule out all of the parameter space. This approach has the same limitations as the first, of always concluding successful replication with a low sample size, and never concluding successful replication with a large sample size.

Fourth, we assess (4) a combined regression analysis. We assessed whether the replication was successful by combining the two studies into a single linear regression analysis. A successful replication is then drawing the same conclusion using the combined data as the original data. This analysis improves the power of the statistical tests due to the increased sample size achieved by combining the two samples.

When considered together, these tests provide insight into whether the replication was successful. One of the primary challenges in assessing whether a replication is successful is accounting for Type II error (i.e. incorrectly accepting the null hypothesis). In the context of replication, this is the probability of incorrectly finding that the replication is successful, when in truth it is not. In this study, the sample size of the replication is constrained by the existing SBO participant pool, which limited our ability to perform a higher-powered test and increases the chance of Type II errors. To account for this, we interpret a failure to reject the null hypothesis (i.e. finding that there is no difference in effect size or hypothesis test result) as a lack of evidence of a difference, rather than evidence that there is no difference. Similarly, confidence intervals tend to be larger when the sample size and statistical power are lower, increasing the likelihood that our replication meets our definition of success. Therefore, it is critical to not over-interpret these results. Rather, this is a first attempt to use data logs for validation. As more data is collected, the strength of replication studies using this approach will improve.

3.6 Analysis

In the analysis that follows, we first reproduce the phishing detection experiment by Canfield et al. [4] to assess whether SBO participants perform differently than Amazon Mechanical Turk [32] participants (*mTurk*). We assess differences between the samples using t-tests (t), Chi-squared tests (χ^2), and 2-sided Mann-Whitney-Wilcoxon (W) tests where appropriate. Given the large number of statistical tests across disparate analyses, we generally use $\alpha = .01$ as a threshold for interpretation, rather than applying separate corrections to groups of tests. We replicate the estimation of the SDT parameters and the linear regression analysis to determine any differences in which factors predict performance. In the regression analysis, with 11 independent tests and $\alpha = .05$, we would expect to find at least one false positive (55% chance). Using $\alpha = .01$ reduces this chance to 11%. However, using $\alpha = .01$ is conservative for Type I errors, but not Type II errors. Therefore, we interpret significance using $\alpha = .05$ for the replication (where Type II error matters most) and $\alpha = .01$ for the remaining analysis (where Type I error matters most).

Second, we assess the experimental measures' construct validity with the Pearson correlation between the SDT parameters and a validated measure of security intentions, the Security Behavior Intentions Scale (SeBIS) [6].

Third, we assess predictive validity by whether the SDT parameters improve the fit of logistic models for predicting observations of negative computer security outcomes for SBO participants (i.e. observations of malicious URLs, files, and software). For each outcome, we construct a logistic regression model comprised of the SDT parameters and other predictors of exposure and risky behavior. This serves to test two hypotheses. We expect users who are more susceptible to phishing on the experimental measure to experience more negative computer security outcomes in real life. Thus, our first hypothesis is:

> *H1: Users who are more susceptible to phishing in the SDT experiment (i.e. are less able to detect and avoid threats) are more likely to visit malicious URLs and have malware and malicious files on their computer.*

We test H1 using a likelihood ratio test, which compares goodness of fit for nested logistic regression models with and without the SDT parameters. The likelihood ratio test is the most efficient test of the null hypothesis that the SDT measures do not increase the likelihood of the data given the SDT measures [15, 16]. The second hypothesis we test is:

> *H2: Users who use their computers more (i.e. have greater exposure) or engage in more risky behavior are more likely to visit malicious URLs as well as have malware and malicious files on their computer.*

We test H2 in the construction of the logistic regression models, following the procedure recommended by Hosmer et al. [16].

4. RESULTS

4.1 Sample

We recruited 132 SBO participants to participate in the phishing detection experiment. Of those, 121 started the survey and 98 finished (= 74% response rate). We excluded 5 participants who sent the SBO less than 7 days of data. The final sample (see SBO Sample in Table 2) represents 44% (= 93/213) of all the SBO participants at that time (All SBO in Table 2). As shown in Table 2, the SBO sample was older, $t(121) = 4.52$, $p < .001$, Cohen's $d = 0.69$, and had a higher proportion of college-educated individuals, $\chi^2(1) = 6.83$, $p = .009$, $\phi = 0.17$, than did the mTurk sample in Canfield et al. [4].[1] There was no difference in gender, $\chi^2(1) = 0.05$, $p = .823$, $\phi = 0.01$. Within the SBO sample, older participants tended to be better educated, in part because some of the younger participants were in college (thus had not finished their educations), $r(92) = 0.37$, $p < .001$. Our SBO sample resembled the wider SBO population on these variables (Table 2).

Table 2. Comparison of mTurk and SBO demographics. The mTurk sample is from Canfield et al. [4].

Variable	mTurk	SBO Sample	All SBO
Female	58%	60%	61%
Bachelors+	45%	63%	58%
Age	32 [19, 59]	41 [19, 81]	46 [19, 87]
N	152	93	213

4.2 Comparison of Experimental Results (Replication)

There was little difference between how much attention the SBO and mTurk participants paid to instructions. Of the 93 SBO participants, 16 failed at least 1 of the 3 attention checks. Users who failed the attention checks were not excluded from the sample, but attention was included as a variable in the regression analysis in order to increase statistical power [30]. There were no significant differences in performance on the attention checks, 17% failed for SBO vs. 10% failed for mTurk, $\chi^2(1) = 2.18$, $p = .14$, $\phi = 0.09$. The median time spent on the introductory phishing information was slightly higher for the mTurk participants, *SBO* = 0.74 minutes ($M = 1.16$, $SD = 1.79$) vs. *mTurk* = 0.95 minutes ($M = 3.17$, $SD = 11.51$), $W = 5018$, $Z = 2.25$, $p = .02$, $r = 0.14$.

However, SBO participants, particularly the older ones, spent more time on the individual email stimuli. The median time to complete the experiment was 47 minutes, including breaks (M = 59 min, SD = 49 min). This estimate excludes seven outliers, participants who appeared to stop working and leave the experiment open on their browser for 19 hours to almost 2 weeks. SBO participants spent more time per email, *SBO* = 0.94 minutes ($M = 1.13$, $SD = 0.72$) vs. *mTurk* = 0.48 minutes ($M = 0.53$, $SD = 0.24$), $W = 11850$, $Z = 8.88$, $p < .001$, $r = 0.57$ in a Mann-Whitney-Wilcoxon test. Within the SBO sample, older participants spent more time per email, $r(92) = 0.46$, $p < .001$.

First, we assess whether the results of the SDT parameter estimation replicate. Since these are point estimates, there are no hypothesis tests to replicate. There was no evidence of significant differences between the mTurk and SBO samples on any SDT parameters, for either the detection or the behavior task, $p > .05$. However, the point estimates differ by 0.12 for detection c, which exceeds our meaningful difference threshold. When comparing the confidence intervals, the replicated point estimate is within the original study's confidence interval for d' and behavior c. For detection c, the replicated point estimate is outside of the original confidence interval, but the confidence intervals still overlap. In general, there is no evidence that the SDT estimates differ between the studies, although the evidence is weakest for detection c. Table 3 shows the mean statistics for the SDT parameters and accuracy (for comparison). Figure 1 shows the distribution of d' and c for each task and sample. There was no evidence of learning over the course of the experiment, as d' and c were equal when calculated separately for the first and second half of the emails. This suggests that the performance parameters estimated in Canfield et al. [4] are not unique to mTurk and can be generalized to the SBO population, which was an older, potentially less tech-savvy group.

We also replicated the regression analysis from Canfield et al. [4] to determine whether there were any differences in the factors that predicted phishing vulnerability for the two samples. Tables 4 and 5 show the results for both samples to compare the results of the hypothesis tests. Figure 2 compares the 95% confidence intervals. In general, the SBO sample's coefficients had larger confidence intervals, due to the lower sample size, but overlap the mTurk coefficients, suggesting no statistically significant differences. The results were largely the same, except for the following three differences.

First, unlike Canfield et al.'s mTurk sample, confidence was not a significant predictor of response bias (c) for the SBO sample. We found no systematic differences in mean confidence between the two samples, $M = 0.86$ ($SD = 0.08$) for SBO and mTurk, $t(181) = 0.04$, $p = .97$, Cohen's $d = 0.01$. Second, age and education are predictors of c in the SBO sample, but were not in the mTurk sample, perhaps due to the higher variance of age and education in the SBO sample. Older participants seemed biased toward identifying emails as phishing (i.e. lower detection c). College-educated participants seemed biased toward identifying emails as legitimate (higher detection c). Third, attention and median time per email were not significant predictors for the SBO sample, perhaps due to reduced variance, as SBO participants were more likely to fail the attention checks and spent more time per email.

As also reported in Tables 4 and 5, the combined analysis is largely consistent with the original Canfield et al. experiment for sensitivity, but there are differences for response bias. Higher attention and higher average confidence predict higher detection sensitivity, consistent with the original Canfield et al. ($p < .01$). None of the predictors are significant for behavior sensitivity, consistent with the original Canfield et al. ($p < .01$). Higher average confidence, lower perceived consequences, and younger individuals tended to have a higher detection response bias, which differs from the original Canfield et al. study ($p < .01$). In the separate analysis, age is significant for the SBO sample but not the mTurk sample and average confidence is significant for the mTurk sample but not the SBO sample. Higher average confidence and lower perceived consequences are associated with a higher behavior response bias, which differs from the original Canfield et al. study ($p < .01$). In the separate analysis, the median time spent per email is significant for the mTurk sample and none of the predictors are significant for the SBO sample.

Table 3. SDT phishing vulnerability parameter estimates for mTurk [4] and SBO samples.

	Detection Task (Yes/No)			Behavior Task (multiple choice)			
	mTurk M (SD) [CI]	SBO M (SD) [CI]		mTurk M (SD) [CI]	SBO M (SD) [CI]		Typical Range
Sensitivity (d’)	0.96 (0.64) [0.86, 1.06]	0.96 (0.66) [0.83, 1.10]	$t(191) = 0.01$, $p = .99$, $d = 0$	0.39 (0.50) [0.31, 0.47]	0.42 (0.52) [0.32, 0.53]	$t(190) = 0.41$, $p = .69$, $d = 0.05$	0 to 4
Response bias (c)	0.32 (0.46) [0.24, 0.39]	0.20 (0.51) [0.10, 0.30]	$t(178) = -1.78$, $p = .08$, $d = 0.24$	-0.54 (0.66) [-0.64, -0.43]	-0.62 (0.57) [-0.74, -0.51]	$t(216) = -1.07$, $p = .29$, $d = 0.14$	-2 to 2
Accuracy	0.67 (0.11) [0.65, 0.69]	0.67 (0.11) [0.65, 0.69]	$t(193) = 0.03$, $p = 0.98$, $d = 0$	0.56 (0.08) [0.55, 0.57]	0.57 (0.09) [0.55, 0.59]	$t(179) = 0.99$, $p = .32$, $d = 0.13$	0 to 1

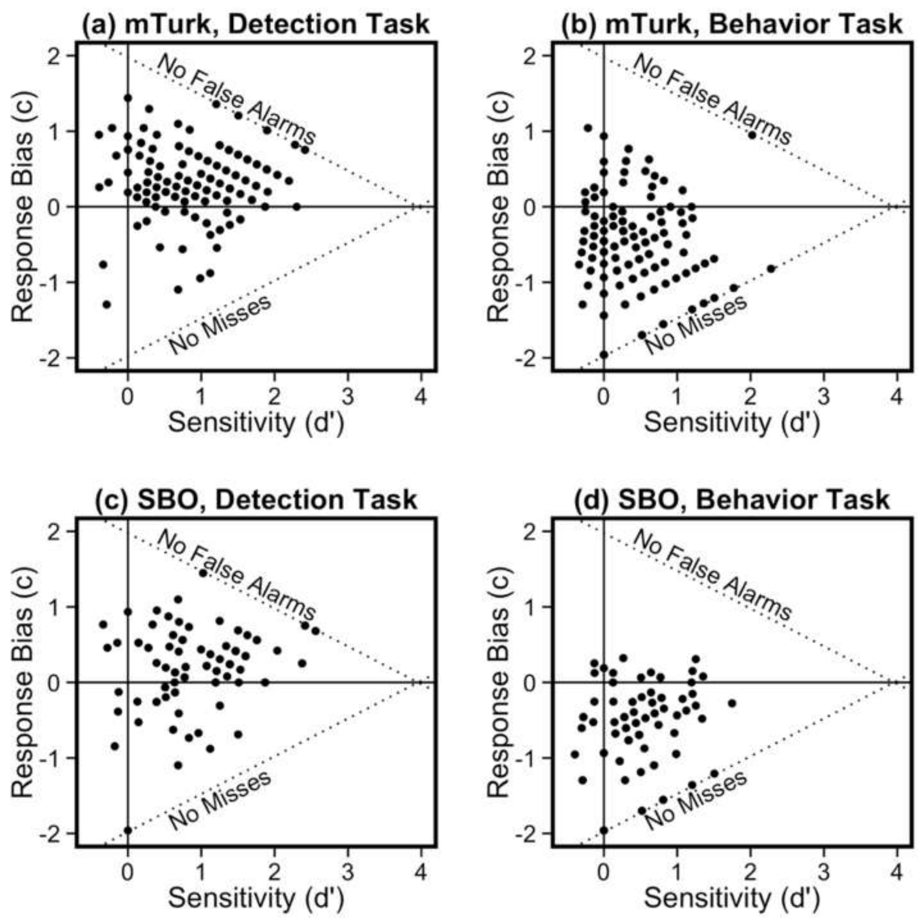


Figure 1. Plot of d’ vs. c for each task and sample. The parameter estimates are bounded by the dotted lines, which represent extreme performance (no false alarms or no misses). There were no significant differences in performance between the mTurk (a, b) [4] and SBO (c, d) samples.

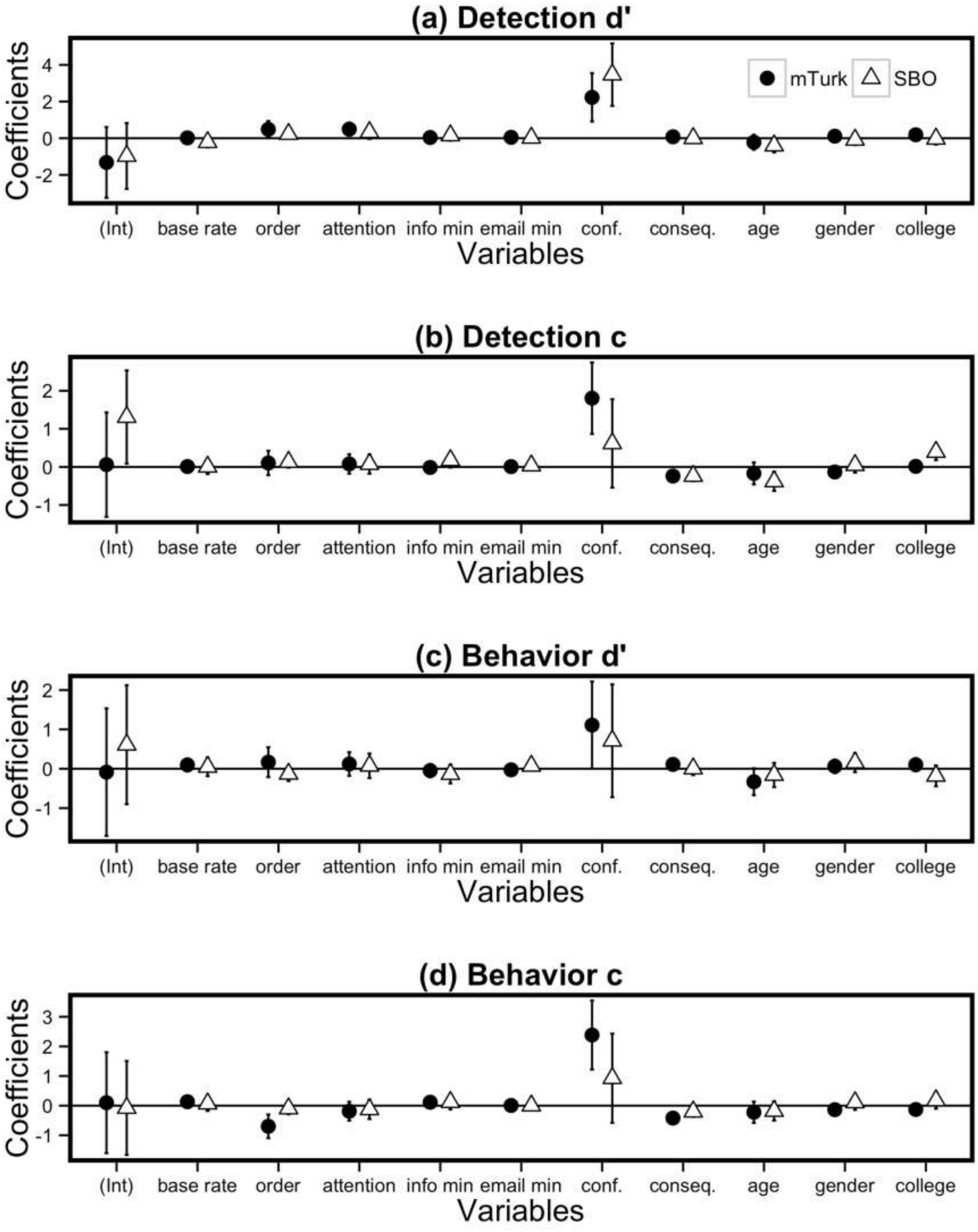


Figure 2. Comparison of regression coefficients with 95% confidence intervals (CI) for (a) detection d', (b) detection c, (c) behavior d', and (d) behavior c.

Table 4. Comparison of linear regression analysis of detection and behavior sensitivity (d') for mTurk [4] and community (SBO) samples. The asterisks indicate statistical significance, where * is p < .05, ** is p < .01, and * is p < .001.**

	Detection Sensitivity (d'$_D$)			Behavior Sensitivity (d'$_B$)		
	mTurk	SBO	Combined	mTurk	SBO	Combined
	B (SE)	B (SE)	B (SE)	B (SE)	B (SE)	B (SE)
Intercept	-1.32 (0.98)	-0.97 (0.92)	-0.96 (0.64)	-0.09 (0.83)	0.61 (0.77)	0.19 (0.54)
Sample (SBO = 1)			-0.04 (0.10)			0.15 (0.08)
Knowledge of base rate	0.02 (0.10)	-0.22 (0.14)	-0.05 (0.08)	0.10 (0.08)	0.05 (0.12)	0.08 (0.07)
Task order (detection = 1)	0.04 (0.10)	0.15 (0.14)	0.09 (0.08)	-0.05 (0.09)	-0.14 (0.12)	-0.06 (0.07)
Attention (pass = 1)	0.49 (0.18)**	0.33 (0.19)	0.40 (0.13)**	0.12 (0.15)	0.07 (0.16)	0.10 (0.11)
log(Phish info time)	0.05 (0.04)	0.02 (0.07)	0.05 (0.03)	-0.03 (0.03)	0.07 (0.06)	0 (0.03)
Median time/email	0.48 (0.23)*	0.23 (0.11)*	0.21 (0.09)*	0.17 (0.19)	-0.13 (0.09)	-0.08 (0.08)
Average confidence	2.23 (0.67)**	3.46 (0.87)***	2.64 (0.51)***	1.11 (0.57)	0.71 (0.73)	0.84 (0.43)
Average perceived consequences	0.08 (0.08)	0 (0.10)	0.07 (0.06)	0.11 (0.06)	0 (0.08)	0.08 (0.05)
log(Age)	-0.22 (0.21)	-0.40 (0.19)*	-0.33 (0.13)*	-0.33 (0.17)	-0.16 (0.16)	-0.26 (0.11)*
Gender (male = 1)	0.11 (0.10)	-0.09 (0.15)	0.04 (0.08)	0.06 (0.09)	0.15 (0.12)	0.10 (0.07)
College (college degree = 1)	0.19 (0.10)	-0.03 (0.16)	0.11 (0.08)	0.10 (0.09)	-0.18 (0.13)	0.02 (0.07)
N	142	84	227	142	84	227
Adjusted R^2	0.16	0.14	0.15	0.05	0.05	0.05
F	3.71***	2.37*	4.63***	1.68	1.40	2.16*

Table 5. Comparison of linear regression analysis of detection and behavior response bias (c) for mTurk [4] and community samples. The asterisks indicate statistical significance, where * is p < .05, ** is p < .01, and * is p < .001.**

	Detection Response Bias (c_D)			Behavior Response Bias (c_B)		
	mTurk	SBO	Combined	mTurk	SBO	Combined
	B (SE)	B (SE)	B (SE)	B (SE)	B (SE)	B (SE)
Intercept	0.06 (0.70)	1.31 (0.62)*	0.81 (0.46)	0.10 (0.87)	-0.08 (0.81)	-0.14 (0.58)
Sample (SBO=1)			-0.12 (0.07)			0.13 (0.09)
Knowledge of base rate	0.01 (0.07)	0 (0.10)	-0.01 (0.06)	0.13 (0.09)	0.07 (0.13)	0.10 (0.07)
Task order (detection=1)	-0.01 (0.07)	0.18 (0.10)	0.01 (0.06)	0.11 (0.09)	0.12 (0.13)	0.08 (0.07)
Attention (pass = 1)	0.08 (0.13)	0.07 (0.13)	0.07 (0.09)	-0.19 (0.16)	-0.13 (0.17)	-0.13 (0.12)
log(Phish info time)	0.01 (0.03)	0.04 (0.05)	0.01 (0.02)	0.01 (0.04)	0 (0.06)	0 (0.03)
Median time/email	0.10 (0.16)	0.13 (0.08)	0.14 (0.06)*	-0.70 (0.20)***	-0.10 (0.10)	-0.17 (0.08)*
Average confidence	1.81 (0.48)***	0.62 (0.59)	1.30 (0.36)***	2.38 (0.59)***	0.93 (0.77)	1.92 (0.47)***
Avg perceived consequences	-0.24 (0.05)***	-0.24 (0.07)***	-0.26 (0.04)***	-0.42 (0.07)***	-0.20 (0.09)*	-0.36 (0.05)***
log(Age)	-0.17 (0.15)	-0.38 (0.13)**	-0.27 (0.09)**	-0.22 (0.18)	-0.18 (0.16)	-0.21 (0.12)
Gender (male=1)	-0.13 (0.07)	0.05 (0.10)	-0.06 (0.06)	-0.14 (0.09)	0.11 (0.13)	-0.05 (0.08)
College (college degree=1)	0.02 (0.07)	0.39 (0.11)***	0.12 (0.06)*	-0.13 (0.09)	0.18 (0.14)	-0.02 (0.08)
N	142	84	227	142	84	227
Adjusted R^2	0.18	0.27	0.21	0.39	0.07	0.25
F	4.16***	4.12***	6.44***	9.85***	1.63	7.81***

4.3 Construct Validity

We assessed construct validity as the correlation between the SDT parameters and the proactive awareness subscale of the SeBIS. One of the four SDT parameters, behavior c (i.e. how suspicious a link must be before the participant chooses not to click on it), was correlated with the SeBIS proactive awareness subscale, $r(83) = -0.29$, $p = .008$. None of the other SDT parameters had a correlation greater than 0.20. Thus, participants who reported looking at the URL before clicking on links (in the SeBIS) were also more cautious in the experimental task (behavior c).

4.4 Predictive Validity

For simplicity's sake, we only report tests of predictive validity for the behavior task, as results for the detection task were similar. Below, we report our analyses separately for each of the four SBO computer security negative outcomes.

4.4.1 Malicious URLs in Browser Data

Browser data were available for 86 of the 93 SBO users. Most used Internet Explorer (66/86 = 77%), followed by Chrome (29/86 = 34%) and Firefox (12/86 = 14%). Some participants used multiple browsers, so the percentages do not sum to 100%. In total, 9 participants (10%) had visited a malicious URL: 2 Internet Explorer users (2/66 = 3%), 4 Chrome users (4/29 = 14%), and 3 Firefox users (3/12 = 25%).

Table 6 shows our univariate analyses [16] for the relationship between each potential predictor and whether users had visited a malicious URL. Among these potential covariates, only domains/day was related to whether participants had visited malicious URLs. Therefore, it was included in the regression model, using a log transformation to normalize the observations.

Users who visited more domains were more likely to have visited a malicious URL. Table 8 shows the regression analysis for the browser data. Log(domains/day) was the only significant predictor. As seen in the likelihood ratio test (reported in the last row of Table 8), users' SDT parameter estimates did not improve the model fit. This indicates that there was no evidence that ability to identify phishing emails in the experiment (as represented by the SDT parameters) was related to whether participants had visited a malicious URL in the browser data.

4.4.2 *Malicious URLs in Network Packet Data*

We also assessed visits to malicious URLs in the network packet data. There was much more network packet data than browser data (Table 6), since a single webpage is assembled from many network packets [17]. For 31 of 93 SBO users (33%), the network packet data indicated that they had visited a malicious URL. Univariate analysis [16] found that total URLs/day, unique URLs/day, and domains/day were related to having visited a malicious URL at least once. We then computed a factor analysis, which revealed that these covariates loaded on one factor, alpha = 0.79. We called this factor *browsing intensity* and used a log transformation to normalize it. We then used that factor score in the regression model and likelihood ratio test reported in Table 8.

The regression analysis shows that users with higher browsing intensity were more likely to have visited a malicious URL in the network packet data. In addition, there was an effect for gender, whereby men were more likely to have visited malicious URLs. This finding emerges after normalizing for exposure (in the regression analysis) and observing no correlation between gender and exposure, $r(90) = .06$, $p = .57$. This suggests that men were either more likely to visit malicious URLs in their browsing or worse at detecting malicious URLs in this sample. More research is needed to understand this result. In the likelihood ratio test, users' SDT parameter estimates did not improve the model fit. Thus, there was no evidence to suggest that performance on the SDT experiment was related to whether participants had visited a malicious URL in the network packet data.

Table 6. Descriptive statistics and factor analysis for the browser and network packet sensor predictors.

	Browser		Network Packet		
	Median	M (SD)	Median	M (SD)	Loading
Days	40	67 (76)	70	85 (63)	NA
Total URLs	22	56 (90)	1,500	2,600 (3,600)	0.73
Unique URLs	9	23 (32)	670	990 (1,000)	1
Domains	5	5.7 (4.4)	42	52 (37)	0.60
% of Total Variance					63%
Cronbach's Alpha					0.80

4.4.3 *Malware*

Most users had the Windows 10 operating system (53/92 = 58%), followed by Windows 8 (22/92 = 24%), Windows 7 (14/92 = 15%), and Windows Vista (3/92 = 3%). 43 of the 92 (47%) users with installed software data had malware. For each operating system, approximately half of the users had malware.

Table 7 shows descriptive statistics for viable software covariates. Univariate analysis [16] revealed that total software and delayed software updates were related to malware. However, the factor analysis found that these variables were only weakly related. When included in the regression model separately, delayed software updates were not a significant predictor, so it was removed from the model. Total software was normalized using a log transformation.

Users who installed more software were more likely to have malware on their machine. As shown in Table 8, this variable predicted malware. Again, the SDT parameter estimates did not improve the model fit. Thus, there was no evidence that performance on the SDT experiment was related to observations of malware on a participant's computer.

4.4.4 *Malicious Files*

Most users (84/93 = 90%) had malicious files on their computer. In the regression model, we used the same predictors as in the malware model, reported in Table 7.

The regression analysis (Table 8) shows that users who had installed more software were significantly more likely to have malicious files on their computer. The SDT parameter estimates did not improve the model fit. Thus, there was no evidence that performance on the SDT experiment was related to observations of malicious files on a participant's computer.

Table 7. Descriptive statistics and factor analysis for the software predictors.

	Median	M (SD)	Loading
Total Software	244	342 (316)	0.44
Delayed Software Updates	2	2 (1)	0.44
% of Total Variance			20%
Cronbach's Alpha			0.33

5. DISCUSSION

In this study, we reproduced Experiment 1 from Canfield et al. [4] in a community sample (SBO). We assessed replicability in terms of the effect sizes, results of the hypothesis tests, confidence intervals, and combined analysis. In general, we found similar distributions of the SDT performance measures as in the mTurk sample, suggesting that there was no evidence of differences in performance between the two samples. However, although the performance of the two samples replicated (as defined in Section 3.5), the regression analysis differed slightly, reflecting the differences between the samples in terms of age and education. This analysis suggests that a higher-powered study with a diverse sample is needed to assess demographic effects. However, the findings about confidence and perceived consequences are fairly consistent, suggesting that they may be useful parameters for future behavioral interventions and predictive metrics.

We found some evidence of construct validity for the experimental behavior task, consistent with it measuring what it claimed. Participants with a greater response bias on the behavior task (c_B), or tendency to treat uncertain emails as phishing, had higher scores on the SeBIS proactive awareness subscale, which elicits self-reports of attention to URLs. This suggests that participants were acting on their computer security intentions in the SDT experiment. The other SDT parameters were not correlated with SeBIS. This suggests that ability (d') is not related

Table 8. Logistic regression models and likelihood ratio test (LRT) for each outcome. The LRT compares the full models shown above with the same models excluding the 2 SDT parameters. The asterisks indicate statistical significance, where * is p < .05, ** is p < .01, and * is p < .001.**

	Malicious URLs (browser)	Malicious URLs (network packet)	Malware	Malicious Files
(Int)	-6.43 (2.14)**	-10.53 (2.83)***	-5.93 (1.71)***	-6.65 (3.71)
Behavior Sensitivity (d'_B)	-0.06 (0.89)	-0.33 (0.55)	-0.09 (0.46)	-1.59 (1.04)
Behavior Response Bias (c_B)	-0.80 (0.74)	0.11 (0.50)	-0.06 (0.44)	-0.90 (1.22)
log(Domains/day)	1.93 (0.77)*			
log(Browsing Intensity)		1.39 (0.38)***		
log(Total Software)			0.99 (0.31)**	2.58 (0.87)**
Age	0.01 (0.03)	-0.03 (0.02)	0 (0.01)	-0.05 (0.03)
Male	0.90 (0.81)	1.47 (0.55)**	0.07 (0.48)	-0.64 (0.94)
College	-0.89 (0.95)	0.16 (0.61)	0.56 (0.53)	-1.29 (1.29)
LRT	$\chi^2(2)$ =1.29, p = 0.5	$\chi^2(2)$ = 0.41, p = 0.8	$\chi^2(2)$ = 0.06, p =1.0	$\chi^2(2)$ = 4.12, p = 0.13

to security intentions. The response bias (c) for the detection task measures participants' tendency to identify emails as phishing or legitimate. Although this could have been related to security intentions, the behavior task better matched the SeBIS scale due to the higher consequences associated with behavior.

We found no evidence of predictive validity for the SDT parameters for any of the four computer security outcomes in the SBO data: browser visits to malicious URLs and network packet data, malware, and malicious files. Thus, we reject H1. However, those four measures were robust enough to be predicted by other observation-based measures, as hypothesized by H2. SBO participants who used their computers more frequently were more likely to have experienced a negative computer security outcome.

We offer four possible reasons why the ability to identify suspicious messages in the laboratory task did not predict the ability to identify similar suspicious messages in the real world:

1. the experimental task does not evoke true phishing behavior,
2. the experimental task evokes true behavior in an environment different from SBO users' (i.e. lack of ecological validity),
3. the SBO measures are confounded by other aspects of users' complex real-world experience, or
4. the SBO data are too noisy to reveal the underlying correlations without much larger samples.

Explanation (1), that the experiment does not evoke actual behavior, seems unlikely, as the results of the experiment are in line with other phishing susceptibility research. For example, participants who perceived worse consequences were more cautious (negative c) [34, 42, 47]. Moreover, performance on the SDT experiment showed expected correlations with other variables, such as better performance being associated with greater security intentions (in the test of construct validity).

Explanation (2), lack of ecological validity for the experiment environment, seems more plausible. One unrepresentative feature of the experimental task is that it has a 50% base rate of phishing emails, much higher than that in everyday life [34]. That higher rate seems likely to have influenced the SDT estimates. In a SDT study of baggage screening, artificially high base rates decreased c (i.e. encouraged participants to be more biased toward identifying items in baggage as suspicious), but had no effect on d' (i.e. people's ability to differentiate between suspicious and benign items in baggage) [46]. Analogous behavior here would have been a greater propensity to treat messages as phishing in the experiment than in life. A second feature of the experimental task is explicitly asking participants to evaluate each email for phishing, thereby priming them to detect attacks. Research by Parsons et al. [33] suggests that explicitly mentioning phishing artificially increases d' but has no effect on c. Together, these studies suggest that our estimates of performance are better than what would be expected in real life. However, there is no obvious reason why these differences should affect users' relative performance. Thus, we would expect users who are good at detecting phishing to perform better on the experiment than users who are bad at detecting phishing. As a result, the correlations across measures should be preserved. In other words, we would not expect users who are bad at detecting phishing in real life to be better at it in an experiment, compared to users who are good at detecting phishing in real life.

Explanation (3), that the complexity of real-world environments (for SBO participants, among others) complicates the relationship between individuals' general propensities (which the SDT metrics attempt to measure) and their actual experiences, is also compelling. As seen here, negative experiences (in the sense of visiting malicious URLs and having malicious files) are strongly related to the amount of exposure (as measured by browsing intensity and total software). Perhaps individuals' exposure to threats overwhelms their ability (d') or propensity (c) to avoid them. Alternatively, the ability to detect phishing emails may not translate to users' ability to avoid attack vectors in general. Thus, the effect of avoiding threats from phishing is washed out by all the other attacks that lead to malware and malicious files on users' computers. Participants' rate of negative experiences may also be related to their systems' protections and their attractiveness as targets for attackers. Systems' vulnerability is partially determined by users (e.g. abilities, knowledge) and partly by others (e.g. browser blacklists, security software). Unfortunately, even with the rich SBO data set, we lacked the complete picture needed to sort out these relationships. The SBO collects data on browser warnings, but there were very few observations. Examining browser warnings would allow observation of the URLs that users attempted to visit, rather than being limited to the successful ones that were not blocked by browser blacklists. In addition, as described in the Methods section, we were unable to measure detections for all security software. Some of those programs, particularly free versions, do not record logs. Others have poor documentation. On those that do provide logs, we observed few detections. Given that security software use did not predict the presence of malware or malicious files and that more malware and malicious files were observed than were detected by security software, one possible explanation is that many SBO users were unable to configure and utilize their security software effectively.

Finally, explanation (4), that the SBO measures are noisy, is to be expected for real-world observations. There were cases where data were missing (e.g. a sensor malfunctioned or was turned off) or ambiguous (e.g. multiple people using the same computer). As a partial check on one potential source of noise, we repeated the analysis after excluding computers with multiple users, but found similar results. If data problems are randomly distributed, then a larger sample might reveal underlying relationships. If they are correlated with individual or system performance, then those interdependencies will need to be understood and unraveled.

Thus, validating predictive measures of phishing vulnerability (including SDT and SeBIS) requires a much more nuanced picture than we currently have of the relationship between individuals' ability, propensities, and experienced outcomes. The predictive validity of any measure could be undermined by proper environmental safeguards or if people realize their vulnerability and restrict their behavior. Once available, a full picture of the SBO data may provide valuable guidance on these possibilities.

5.1 Limitations

This study had several notable limitations. First, it was limited to Windows users. The depth and breadth of SBO data collection requires custom software tailored to each OS. Due to resource constraints, the SBO is limited to Windows, the most common OS [9]. In the original mTurk sample [4], 84% of participants used Windows and performed similarly to other OS users.

Second, although this study evaluates the generalizability of an existing method, it leaves some aspects of generalizability open to further study. Although the mTurk and SBO samples differed in some ways (Table 2), extension to other populations would be valuable. One within-sample result bearing further attention is the observation that men were more susceptible to phishing.

Third, both the experimental task and the SBO study whether individuals visit a phishing website. That leaves open the question of when they share personal information once there. As noted, even the simpler outcome of such visits was difficult to measure in the SBO. We were limited by the data available in the Google Safe Browsing, ShouldIRemoveIt.com, and VirusTotal datasets. Thus, we missed attacks absent in these databases. In addition, we observed more negative computer security outcomes related to software (47% had malware and 90% had malicious files) than to browsing (10% in browser data and 33% in network packet data). This lower rate may reflect limits to the lists of malicious URLs, which change over time. For example, a legitimate site may be compromised and only briefly appear on the Google Safe Browsing blacklist. Finally, some SBO data were missing for technical reasons, which reduced our ability to observe negative outcomes and correlate them with other measures.

5.2 Recommendations

Given the novelty of using data logs like those collected by the SBO to validate performance tests like those collected in Canfield et al., we provide recommendations for future work:

1. To the extent possible, use behavioral outcomes that are (a) as directly related to the outcome of interest as possible and (b) rely on human ability without intervening technology. For example, measure attempts to visit malicious URLs (via browser warning data), rather than actual visits, to distinguish human ability from browser blacklist effectiveness. When possible, use security software detections of malware and malicious files to assess attempts to download malicious files. Technical constraints and lack of observations limited our ability to use these outcomes.
2. Triangulate between multiple data sources (e.g. assessing both browser and network packet data), with an understanding of their respective strengths and weaknesses. For example, there are more network packet data, but browser data better reflect the URLs that users choose to visit. Beyond the analysis presented here, it may be possible to crosscheck events such as security software scanning with observed active processes on the machine.
3. Consider the temporal sequence of events, such as how periods without security software protection affect the risk of acquiring malicious files.

6. CONCLUSION

We assessed the validity of the SDT measures proposed by Canfield et al. [4] in three ways: (a) replicating their mTurk SDT experiment with SBO participants, (b) assessing construct validity via correlation with the SeBIS proactive awareness subscale, and (c) evaluating predictive validity using negative outcomes observed in the SBO data. Our results suggest (a) that the findings from Canfield et al. [4] generalize to the SBO population and (b) the SDT measures have construct validity, given the correlation between participants' self-reported tendency to look at URLs before clicking links (in the SeBIS) and their caution in clicking links in the SDT study (behavior c). However, we found (c) no evidence of predictive validity, as the SDT measures did not predict negative computer security outcomes observed in the SBO.

One of the primary challenges for this analysis was differentiating between people's ability to protect themselves (by knowing which URLs to avoid) and technical safeguards (such as browser blacklists). Future research, addressing this complication, will offer opportunities for laboratory and observational measures to complement one another in understanding the security ecosystem.

7. ACKNOWLEDGEMENTS

We thank the SBO PIs: Alessandro Acquisti, Nicolas Christin, Lorrie Cranor, Serge Egelman, and Rahul Telang, for providing access to the SBO data. In addition, we thank Lorrie Cranor and Serge Egelman for comments on an earlier draft. We also thank Rick Wash for helpful comments. The Security Behavior Observatory was partially funded by the NSA Science of Security Lablet at Carnegie Mellon University (contract #H9823014C0140); the National Science Foundation, Grant CNS-1012763 (Nudging Users Towards Privacy); and the Hewlett Foundation, through the Center for Long-Term Cybersecurity (CLTC) at the University of California, Berkeley.

8. REFERENCES

[1] Anderson, C. J., Bahnik, S., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R. … Zuni, K. (2016). "Response to Comment on 'Estimating the reproducibility of psychological science,'" *Science* 351(6277), 1037-1039. DOI: 10.1126/science.aad9163

[2] Appel, M. (2012). "Are heavy users of computer games and social media more computer literate?" *Computers & Education* 59(4), 1339–1349. DOI: http://doi.org/10.1016/j.compedu.2012.06.004

[3] Camp, L. J. (2009). "Mental models of privacy and security," *IEEE Technology and Society Magazine* 28(3), 37–46. DOI: http://doi.org/10.1109/MTS.2009.934142

[4] Canfield, C. I., Fischhoff, B., & Davis, A. (2016). "Quantifying Phishing Susceptibility for Detection and Behavior Decisions," *Human Factors* 58(8), 1158-1172. DOI: http://doi.org/10.1177/0018720816665025

[5] Cronbach, L. J. & Meehl, P. E. (1955). "Construct validity in psychological tests," *Psychological Bulletin* 52(4), 281-302. http://psychclassics.yorku.ca/Cronbach/construct.htm.

[6] Egelman, S. & Peer, E. (2015). "Scaling the Security Wall," In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: CHI '15*, 2873–2882. DOI: http://doi.org/10.1145/2702123.2702249

[7] Egelman, S., Harbach, M., & Peer, E. (2016). "Behavior Ever Follows Intention?" In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: CHI '16*,1-5. DOI: http://doi.org/10.1145/2858036.2858265

[8] Etz, A. & Vandekerckhove, J. (2016). "A Bayesian Perspective on the Reproducibility Project: Psychology," *PLOS ONE*. https://doi.org/10.1371/journal.pone.0149794

[9] Forget, A., Komanduri, S., Acquisti, A., Christin, N., Cranor, L. F., & Telang, R. (2014). "Security Behavior Observatory: Infrastructure for long-term monitoring of client machines," Technical Report CMU-CyLab-14-009, CyLab, Carnegie Mellon University, Pittsburgh, PA. https://www.cylab.cmu.edu/files/pdfs/tech_reports/CMUCyLab14009.pdf

[10] Forget, A., Pearman, S., Thomas, J., Acquisti, A., Christin, N., Cranor, L. F., Egelman, S., Harbach, M., & Telang, R. (2016). "Do or Do Not, There Is No Try: User Engagement May Not Improve Security Outcomes," In *Proceedings of the Symposium on Usable Privacy and Security: SOUPS '16*, 97–111.

[11] Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). "Comment on 'Estimating the reproducibility of psychological science,'" *Science* 351(6277), 1037-1039. DOI: 10.1126/science.aad7243

[12] Global Stats. (2016). "Top 7 OSs," http://gs.statcounter.com/#desktop-os-ww-monthly-201610-201610-bar

[13] Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). "What does research reproducibility mean?" *Science Translational Medicine* 8(341), 1-7. DOI: 10.1126/scitranslmed.aaf5027

[14] Google. (2016). "Google Safe Browsing APIs (v4)". https://developers.google.com/safe-browsing/v4/

[15] Hauck, W. W. & Donner, A. (1977). "Wald's Test as Applied to Hypotheses in Logit Analysis," *Journal of the American Statistical Association* 72(360), 851-853.

[16] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.), Hoboken, NJ: John Wiley & Sons.

[17] HTTP Archive. (2016). "Trends." Retrieved on June 29, 2016 from http://httparchive.org/trends.php#bytesTotal&reqTotal**.**

[18] Ion, I., Reeder, R., & Consolvo, S. (2015). "'...no one can hack my mind': Comparing Expert and Non-Expert Security Practices," In *Proceedings of the Symposium on Usable Privacy and Security: SOUPS '15,* 327–346.

[19] Jakobsson, M. & Ratkiewicz, J. "Designing ethical phishing experiments: a study of (rot13) ronl query features," In *Proceedings of the 15th International Conference on World Wide Web: WWW '06*, 513–522.

[20] Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L. F., & Hong, J. (2010). "Teaching Johnny not to fall for phish," *ACM Transactions on Internet Technology* 10(2), 1–31.

[21] Lalonde Levesque, F., Nsiempba, J., Fernandez, J. M., Chiasson, S. & Somayaji, A. (2013). "A clinical study of risk factors related to malware infections," In *Proceedings of the ACM SIGSAC Conference on Computer & Communications Security: CCS '13*, 97–108. DOI: http://doi.org/10.1145/2508859.2516747

[22] Leukfeldt, E. R. (2015). "Comparing victims of phishing and malware attacks," *International Journal of advanced studies in Computer Science and Engineering* 5(5), 26-32.

[23] Long, L. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*, Thousand Oaks, CA: Sage Publications.

[24] Macmillan, N. A. & Creelman, D. C. (2004). *Detection Theory: A User's Guide*, New York, NY: Psychology Press.

[25] Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., … Van Der Laan, M. (2014). "Promoting transparency in social science research," *Science* 343, 30-31.

[26] National Institute for Standards and Technology (NIST). (2012). "Guide for Conducting Risk Assessments," NIST Special Publication 800-30, Washington, DC, USA. http://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-30r1.pdf

[27] Nichols, A. L. & Maner, J. K. (2008). "The good-subject effect: Investigating participant demand characteristics," *Journal of General Psychology* 135(2), 151-166.

[28] Nosek, B. A. & Lakens, D. (2014). "Registered Reports: A Method to Increase the Credibility of Published Results," *Social Psychology* 45, 137-141.

[29] Open Science Collaboration. (2015). "Estimating the reproducibility of psychological science," *Science* 349(6251), 943-952. DOI: 10.1126/science.aac4716

[30] Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). "Instructional manipulation checks: Detecting satisficing to increase statistical power," *Journal of Experimental Social Psychology* 45(4), 867–872. DOI: http://doi.org/10.1016/j.jesp.2009.03.009

[31] Orne, M. T. (1962). "On The Social Psychology of the Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications," *American Psychologist* 17(11), 776–783.

[32] Paolacci, G., Chandler, J. & Ipeirotis, P. G. (2010). "Running experiments on Amazon Mechanical Turk," *Judgement and Decision Making* 5(5), 411–419.

[33] Parsons, K., McCormac, A., Pattinson, M., Butavicius, M., & Jerram, C. (2015). “The design of phishing studies: Challenges for researchers,” *Computers & Security* 52, 194-206. DOI: http://doi.org/10.1016/j.cose.2015.02.008
[34] Pattinson, M., Jerram, C., Parsons, K., McCormac, A. & Butavicius, M. (2012). “Why do some people manage phishing e-mails better than others?” *Information Management & Computer Security* 20(1), 18–28.
[35] Schwartz, D., Fischhoff, B., Krishnamurti, T. & Sowell, F. (2013). “The Hawthorne Effect and energy awareness,” *PNAS*, 110(38), 15242-15246.
[36] Sheng, S., Holbrook, M. B., Kumaraguru, P., Cranor, L. F. & Downs, J. (2010). “Who Falls for Phish? A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions,” In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: CHI ’10*, 1-10.
[37] Sheng, S., Kumaraguru, P., Acquisti, A., Cranor, L. F., & Hong, J. (2009). “Improving phishing countermeasures: An analysis of expert interviews,” In *Proceedings of the 4th APWG eCrime Researchers Summit*.
[38] Sotirakopoulos, A., Hawkey, K., & Beznosov, K. (2011). “On the Challenges in Usable Security Lab Studies: Lessons Learned from Replicating a Study on SSL Warnings,” In *Proceedings of the Symposium on Usable Privacy and Security: SOUPS ‘11*.
[39] Symantec Corporation. (2017). *Internet Security Threat Report*. https://www.symantec.com/security-center/threat-report
[40] Verizon. (2017). *2017 Data Breach Investigations Report*. Retrieved from http://www.verizonenterprise.com/verizon-insights-lab/dbir/2017/
[41] Vishwanath, A., Herath, T., Chen, R., Wang, J., Rao, H. R. (2011). “Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model,” *Decision Support Systems* 51(3), 576–586. DOI: http://doi.org/10.1016/j.dss.2011.03.002
[42] Wang, J., Herath, T., Chen, R., Vishwanath, A., & Rao, H. R. (2012). “Phishing susceptibility: An investigation into the processing of a targeted spear phishing email,” *IEEE Transactions on Professional Communication* 55(4), 345–362. DOI: http://doi.org/10.1109/TPC.2012.2208392
[43] Wash, R. (2010). “Folk Models of Home Computer Security,” In *Proceedings of the Symposium on Usable Privacy and Security: SOUPS ‘10*.
[44] Wash, R. & Rader, E. (2015). “Too Much Knowledge? Security Beliefs and Protective Behaviors Among United States Internet Users,” In *Proceedings of the Symposium on Usable Privacy and Security: SOUPS ‘15*, 309-325.
[45] Wasserstein, R. L. & Lazar, N. A. (2016). “The ASA’s statement on p-values: context, process, and purpose,” *The American Statistician* 70(2), 129-133.
[46] Wolfe, J. M., Horowitz, T. S., Van Wert, M. J., Kenner, N. M., Place, S. S. & Kibbi, N. (2007). “Low target prevalence is a stubborn source of errors in visual search tasks,” *Journal of Experimental Psychology: General* 136(4), 623–638. DOI: http://doi.org/10.1037/0096-3445.136.4.623
[47] Wright, R. T. & Marett, K. (2010). “The Influence of Experiential and Dispositional Factors in Phishing: An Empirical Investigation of the Deceived,” *Journal of Management Information Systems* 27(1), 273–303.

APPENDIX

A. Open Data

The data and code for this paper are available at https://osf.io/6dknx/.

Can We Fight Social Engineering Attacks By Social Means? Assessing Social Salience as a Means to Improve Phish Detection

James Nicholson
PaCT Lab
Northumbria University
Newcastle upon Tyne, UK
james.nicholson@northumbria.ac.uk

Lynne Coventry
PaCT Lab
Northumbria University
Newcastle upon Tyne, UK
lynne.coventry@northumbria.ac.uk

Pam Briggs
PaCT Lab
Northumbria University
Newcastle upon Tyne, UK
p.briggs@northumbria.ac.uk

ABSTRACT

Phishing continues to be a problem for both individuals and organisations, with billions of dollars lost every year. We propose the use of nudges – more specifically social saliency nudges – that aim to highlight important information to the user when evaluating emails. We used Signal Detection Theory to assess the effects of both sender saliency (highlighting important fields from the sender) and receiver saliency (showing numbers of other users in receipt of the same email). Sender saliency improved phish detection but did not introduce any unwanted response bias. Users were asked to rate their confidence in their own judgements and these confidence scores were poorly calibrated with actual performance, particularly for phishing (as opposed to genuine) emails. We also examined the role of impulsive behaviour on phish detection, concluding that those who score highly on dysfunctional impulsivity are less likely to detect the presence of phishing emails.

1. INTRODUCTION

Phishing is a highly prevalent form of social engineering where an attacker steals sensitive information by sending fraudulent emails that purport to be from a trustworthy source. Over time, phishing attacks have become both socially and contextually smarter, with the result that phishing continues to be a growing problem for organisations and individuals. In the best-case scenario, phishing results in lost productivity due to users deliberating over the authenticity of the email, but in the worst-case scenario individuals and businesses can suffer serious security, financial and/or reputation loss due to stolen credentials or leaked information.

A large number of people fall for these phishing emails within experimental studies [25, 31, 38]. For example, McAfee's Phishing Quiz [31] found that 80% of respondents (employees) fell for at least one phishing email – an alarmingly high percentage. A recent "in the wild" study showed that users do not only follow the link, they go on to provide their credentials to the website. This study, by Bursztein et al. [6], examined the effectiveness of phishing websites by analysing internet traffic through Google, and found that the most successful phishing web page resulted in 45% of page views converting into captured user credentials. However, not all webpage visits successfully converted to captured credentials, while an average conversion rate of 14% was found across all the websites they looked at. To deal with this issue, researchers have focused on two core strategies: either improving the filtering algorithms that can reduce the number of phishing emails that make it into users' inboxes (e.g. [3, 9]) or developing interventions, mainly training and education, that help users identify fraudulent emails (e.g. [40]). Despite these efforts, both individuals and organisations continue to fall for phishing scams and billions of dollars are lost every year – the Monthly Online Fraud Report for January 2015 estimates losses of over $4.5 billion for 2014 [37].

In the current study, we focus on the second of these strategies, exploring interventions that might support the user in the detection of fraudulent emails. In particular, we wanted to explore the effect of making the broader social context of the email more salient. We did this firstly by highlighting the name of the sender along with the time the email was sent, recognising that genuine emails are typically exchanged during certain social or business hours; and secondly, by highlighting the number of people in an organisation or network that received that same email, recognising that genuine emails are targeted at specific individuals or groups, while phishing emails are more socially indiscriminate.

2. Background Research

As we have noted, attempts to deal with the phishing problem embrace both technical and human-centric solutions. Technical solutions have generally focused on identifying suspicious websites, for example using browser plugins or identifying characteristic elements of a phishing email, e.g. [16]. Filtering algorithms can also bring improvements, e.g., [3, 9], however such phishing tools are not always accurate – some phish are missed and some genuine items are flagged as phish, i.e. there are problems with false positives and negatives [50].

The human-centric solutions typically fall into one of three categories involving (i) educational or training interventions; (ii) new designs and visualisations that can help 'nudge' users to make better decisions and (iii) work that considers individual differences in decision-making. Our work primarily addresses the latter two categories, but we will briefly consider some of the educational initiatives below.

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12 -- 14, 2017, Santa Clara, California.

2.1 Campaigns and Educational Interventions

Users are unlikely to take effective action against phishing attacks unless they are both aware of the risks inherent in online communication and are also knowledgeable about the specific threats posed by dubious emails. Indeed, researchers have shown that the perceived risk of cybercrime can moderate users' willingness to take risks in a variety of online environments [36] and that the ability to evaluate deceptive cues was a major factor in online protection [22]. Further, users' cybersecurity (i.e., phishing) knowledge is positively related to their attitude and intention toward adopting and using cybersecurity (anti-phishing) solutions [46]. It is unsurprising, then, that a number of educational interventions designed to improve user understanding of risk and knowledge of how to mitigate risk have been developed. These interventions adopt a wide range of different training techniques that can include embedded training systems [28] motivational cartoons [29] and even games that raise awareness and train users for future encounters [41]. However, they have been met with limited success. Users start with a very poor awareness of their vulnerability to being phished [23, 46] and may ignore the training altogether [8]. Added to that, the phishing emails become more sophisticated year on year - to the point where even security experts are unable to determine whether the item is genuine or not [18].

2.2 Behavioural Interventions

When seeking to influence user behaviour, we must be mindful that people are reluctant to spend much time and effort engaging in protective privacy or security measures [24, 39]. Many cybersecurity interventions are unproductive and unhelpful in the sense that they take time and effort away from the users' primary task. This productivity argument is important, as employers often do not appreciate how much time is lost due to staff deliberating the legitimacy of emails. Of course, the costs to organisations can be much worse when employees get this wrong and when companies are then laid open to serious cybersecurity threats and can incur significant financial and/or reputation loss.

However, we should also be aware of the vulnerabilities exhibited by users during the regular processes of communicating by email. Ferreira et al., [20] note that the principle of Liking, Similarity & Deception (LSD) rules in this context – as people simply tend to believe in what others do or say as a default, unless they have good reason to suspect something is really wrong or they find a particular behaviour is completely unexpected. With this in mind, many researchers have turned to the principles derived from behavioural economics in order to design a range of seamless cybersecurity "nudges" (see [43]) or visualisations [10, 11] that help move the user away from this default position, so that they make better choices, but choices that do not require too much additional effort on their part. Behavioural nudges are already popular in the privacy field, with successful examples being found in relation to reduced Facebook sharing [47] and improved smartphone privacy settings [1]. They are also becoming popular as cybersecurity interventions, e.g. in relation to the risks associated with the selection and installation of apps on mobile phones [10]. For phishing, the existing interventions typically seek to make the trustworthiness of the linked webpage more salient within the web browser. For example, Chou et al. [12] proposed SpoofGuard, a toolbar that gave pages a *Spoof Score* to help the user evaluate the likelihood that the page is not genuine. This score is based on a URL check (whether URL appears to be genuine), an image check that includes logos (e.g. detecting that an eBay logo does not sit on a non-eBay.com domain), a link check (check that all links in the page point to the current or same domain), and a password check (if page requires password, then more scrutiny is needed). An experimental system called CATINA [51] employed such an approach to obtain a 97% accuracy rate in recognising the phishing websites it examined, with a 6% false positive rate. However, these technical approaches rely on a page being reported as a phishing site before they can be used. Other interventions have explored the effectiveness of browser warnings, including toolbars. This work has generally found warnings and toolbars ineffective (e.g. [48]) – in part due to the user ignoring them. However, further work exploring the design of the phishing warnings on browsers found that active warnings – those requiring an action from the user to be dismissed – were clearly more effective than passive warnings [17].

Other behavioural interventions have focused upon email attachments, which pose a known security problem [14]. Polymorphic Dialogs have been proposed for opening email attachments where, for example, the order of the options might change regularly in order to prevent habituation (or automatic skipping), and a timer can be introduced that forces the user to study and evaluate other options [5]. However, unsurprisingly, such interventions can significantly increase the time taken to complete simple tasks – again, resulting in unacceptable productivity costs for the end user [2].

There have not been many interventions to support the user in detecting the phishing emails themselves within the email client. The aforementioned Polymorphic Dialogue [5] is an example of an intervention built into the email client to deal with attachments, while PhishDuck [49] is another example of a client-based extension designed to deal with phishing links. When a suspicious link is clicked by users, PhishDuck displays a popup asking for confirmation of the action, and presents a suggestion that they may have intended to use a different link (e.g. paypal.com instead of paipal.com). A user study found that participants using the extension followed significantly less phishing links than those using the default email client warning message.

Some email providers and clients will present warnings to users when discrepancies are detected. For example, Gmail displays a banner warning on the top of a message if the email claims to be from a Gmail address but has not been authenticated as such [44] and while this can be a very useful indicator, it only applies to emails from the same domain. The Mozilla Thunderbird email client also displays a banner warning at the top of the email message when (internal) discrepancies are identified [45], but also incorporates the use of a pop up warning requiring the user to click on a continue button if they click on any links within the message. Once again, this fall-back system relies on the automated detection of features within the message that earlier spoofed the spam filter. Finally, it is possible to set up Microsoft Outlook so that users are not able to click on links within emails, but must instead copy and paste (or retype) the URL directly into the web browser. However, this does not attempt to assess if the email is a potential phish and may result in non-discriminatory behaviour from the user to minimise productivity disruption.

2.3 Individual Differences in Susceptibility to Attacks

We already know something about the kinds of people likely to be most vulnerable in a phishing attack. For example, females are more prone than males to misclassifying phishing emails as genuine [25, 26, 29, 40]. Halevi et al. [23] found a relationship between neuroticism and susceptibility to phishing attacks and various work has found that extroverts, more trusting individuals, and those open to new experiences were more vulnerable to phishing attacks [25, 33]. In contrast, Pattinson et al. [35] reported that extraverts and individuals scoring high for openness managed phishing emails better, which they acknowledge as a counter-intuitive finding, but also reported a marginal effect of impulsivity, with those scoring high for impulsivity showing greater susceptibility to phishing attacks, while Modic and Lea [33] hint at an effect of impulsivity by reporting that Premeditation (an item of their impulsivity scale) was the best predictor for scam response rate in their scam compliance survey. Finally, in a recent study of attitudes and behaviours online, Riek et al. [36] have also found an interesting relationship between user confidence, risk perception and the use of online services. Specifically, more confident users have a higher chance of becoming victimized, although they are also more able to identify cybercriminal attacks. This is in contrast with other work in phishing where a positive relationship has been reported between performance (identification) and confidence [7].

In the current study, we have tried to explore nudges that can alert the user to the possible presence of a phishing email. These are simple visual cues that build upon the social premise of a phishing attack – wherein a user is socially engineered to believe that the email comes from a genuine source (e.g. because the sender is known or the content of the email seems appropriate). However, we go further in providing cues that make the social context of the sender more salient (highlighting the name and address of the sender, and the time the email was sent) and the social context of the recipient more salient (highlighting the number of other recipients of that email). We loosely based our two nudges on existing work from other security and privacy contexts, notably installation dialogues that highlight the vendor's name [4] for the former and audience saliency from social media work [47] for the latter (see Section 3.1 for full details).

We hypothesise that each of these should improve phish detection, but we also explore individual differences in user susceptibility to phishing emails, by measuring both functional and dysfunctional impulsivity [13] and user confidence in their own cybersecurity decision-making.

3. Methodology

In order to determine the effectiveness of the two nudges, we set up an online experiment via Amazon's Mechanical Turk where participants were asked to view 18 emails (6 phishing, 12 real) and decide whether each email represented a genuine message or a phishing message. The emails were designed by the researchers but were modelled on real messages received within the previous 3 months. The phishing emails, specifically, were faithful reproductions of emails that had been problematic (as reported by the I.T. department) within the university during that time period.

3.1 Design

The study had a 2 x 2 independent measures design. The first factor, *sender saliency*, was created by highlighting sender features on the email that included name, email address and the time the email was received. This factor had two levels (highlighted, not highlighted). It was chosen, in part, to exploit the social nature of a phishing attack where senders may seem familiar [16] but in all likelihood, the normal "social hours" of that sender would be understood (e.g. it would be unusual to receive an email from a colleague or from a local organisation at 1am). Although the name of the sender can be spoofed, it is common of phishing emails to contain discrepancies between the name of the sender and the original email address. In essence, the sender saliency nudge also aimed to expose any discrepancies in the address field of the emails thereby reducing the likelihood that users would be lured into a false sense of trust. This nudge was modelled on similar security work on installation dialogues showing that highlighting the vendor's details to direct users' attention to potential discrepancies led to more secure behaviour [4]. The sender saliency nudge could be easily deployed in an organisation or to individual users through an email client plug in or using a browser extension.

The second factor, *receiver saliency* was created by informing the user of the number of people within their organisation that also received a version of the email. Again, there were two levels of this factor (receiver information present or absent). This factor was designed to exploit the social context of emails, in that genuine emails are constructed for a particular audience or individual, whereas a spear phishing email from a compromised account may be sent to multiple unrelated recipients. Whilst we recognise that mass emails from popular services (e.g. PayPal) may be sent to multiple recipients, the content or *lure* often appears to be highly personal "Ms x, your account may have been compromised, so please click here to change your password"). If a user is alerted to the fact that a seemingly personal message has been sent to many colleagues, they may question the validity of that message. The converse may also be true – i.e. if they receive a message that should, by its nature, have been distributed to whole organisation (e.g. using a standard mailing list) and yet they are the sole recipient, then again, they may re-evaluate the legitimacy of that message. This nudge was loosely based on the Picture Nudge [47] on Facebook demonstrating that unintended information disclosure could be minimised by alerting the users to the post's target audience. In our case, we have reversed the paradigm where the user instead gets a visual measure of the message's intended audience. The likely deployment of the receiver saliency nudge would be in an organisation where email data can be easily collected to inform the nudge's numerical output.

3.2 Participants

A Human Intelligence Task (HIT) was posted on Amazon's Mechanical Turk (MTurk) stating that we were looking for users willing to help out with an email-sorting task. Participants were given a flat fee of $0.45 for completing the task which had an average completion time of 10 minutes, mirroring the payment structure of other studies at the time. The inclusion criteria for taking part in the study were a minimum age of 18, a good level of English, and a Number of HITs Approved greater than or equal to 50 (for quality purposes). Participants on mobile devices were excluded from participating to control the viewing experience of the emails.

We set recruitment targets based on an *a priori* power analysis suggesting 279 participants for a medium effect, with a final sample of 281 participants then completing the task to the required standard (see Table 1 for details). No attention checks were employed in the experiment, but the data provided was inspected for validity in terms of the time spent on the task: The work from workers who spent two or less seconds on average per email was rejected and new workers were found to complete the study. Participants were randomly allocated to one of the four groups (sender salience cue present/absent; receiver salience cue present/absent).

Table 1: Participant demographics (F=Female; M=Male; U=Undisclosed).

Nudge	N	Mean Age	F	M	U
None (Control)	65	34.5	29	34	2
Sender	64	35.7	31	32	1
Receiver	79	31.7	31	45	3
Combined	73	32.3	20	52	1
Total	281	33.6	111	163	7

3.3 Materials

The 18 emails were presented to participants as static images, with 6 designed as the target phishing emails and the remaining 12 as genuine (see Appendix A.1). Phishing emails were (loosely) matched with genuine emails in terms of the time of day they were received and the percentages of colleagues flagged as also receiving the messages (receiver saliency nudge). This was done by matching six of the genuine emails with the features of the phishing emails (i.e. similar time of day they were received) while the remaining 6 were chosen to reflect the overall established patterns of that set (e.g. most emails received during working hours). Note that this approach is rather conservative, in that we are deliberately reducing the simple effectiveness of our time of day cue as a signal of whether emails are genuine or not, but we are operating on the assumption that some genuine emails may reasonably be received at night (e.g. emails from another continent) and that by alerting users to time sent, we are encouraging them to check the other aspects of the email more carefully.

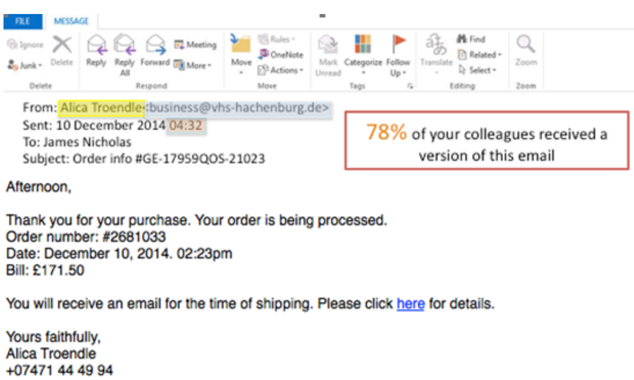


Figure 1: Example phishing email showing both nudges – sender (highlights left) and receiver (box right).

All messages were placed under an image of the Microsoft Outlook Ribbon bar (see Figure 1) to provide a frame of reference to participants. The "to" field in each email was edited to show *James Nicholas* as the receiver and any personal or identifiable information within the email body was edited with generic information.

The emails chosen for this study covered a basic range of possible senders and were matched across phishing and genuine messages: emails from well-known providers (e.g. Amazon, PayPal, eBay), emails from smaller organisations (Spotify, Eversure) and emails from individuals. We chose emails that contained links to websites as these are the most common type of phishing attack by volume [42] although in practice the nudges should work in the same manner with emails containing attachments.

The website was hosted on our own server but the recruitment was facilitated through Amazon's Mechanical Turk.

We note that the phishing emails, modelled on problematic phishing emails received within the university, were designed to always present a cue as to their authenticity to overcome the limitation of having no active links: when the sender metadata (to/from/subject) did not show a clear discrepancy to allow an informed choice, the links in the body of the message were not masked or obscured (similar to previous work [7]).

3.4 Measures

The main dependent variable was whether the user classified each email as either genuine or a phish. This was a binary decision, but the time taken to make a decision (in seconds) on each email was also recorded, starting when the page loaded and concluding when the radio button for the decision was pressed. Finally, participants were asked to rate how confident they were with their own classification of the email as genuine or phish, using a drop-down menu with options ranging from 0% to 100% confident in increments of 10%.

In addition, participants were asked to complete an impulsiveness personality questionnaire at the end. Impulsiveness has been linked with susceptibility to phishing emails in previous work (e.g. [28, 35]). Despite weak associations, the trends reported are interesting enough to warrant further exploration of this aspect of personality in our study. We used a reduced version of Dickman's Impulsivity Inventory [13] and the final scale consisted of 6 items measuring functional impulsivity (acting without much forethought, to maximise efficiency), with an internal reliability of 0.670 (Cronbach's Alpha) and 6 dysfunctional impulsivity items (acting without much forethought, but with undesirable consequences), with an internal reliability of 0.856.

3.5 Procedure

The experiment was initially framed as an email-sorting task on the MTurk HIT, but once participants clicked through to the homepage of the study, they were given more specific instructions telling them they would be required to identify phishing emails. This initial deception was put in place to prevent the recruitment of individuals only interested in computer security. Once on the website, they were randomly assigned to one of the four experimental groups and given the task instructions: they must look through 18 emails that were received by a person called James Nicholas and classify the message as either a genuine email or as a phishing email. After each decision, participants then provided a confidence score for their decision and progressed to the next message. Once all messages had been classified they were thanked and given a code to enter on the Mechanical Turk page. Participants received their payment once their work was reviewed by the research team.

4. Results

4.1 Scoring

The absolute user judgement of genuine/phish was scored in terms of classical signal detection theory, i.e. as a hit, a miss, a true negative or a false positive. In our task, *hit rate* refers to phishing emails that were correctly identified as phishing emails. *False positive rate* (or false alarms) refers to genuine emails that were incorrectly identified as phishing. Signal detection theory was developed to determine the sensitivity of a participant to the presence of a target (phishing emails) against a background of noise (genuine emails). The discriminability index *d'* is a statistic used in signal detection that provides the separation between the means of the signal and the noise distributions in units of standard deviation of noise distributions. *d'* was calculated using the equation:

$$d' = Z(\text{phish hit rate}) - Z(\text{false positive rate})$$

Bias in decision-making (i.e. whether users tend to classify things as phish or as genuine, irrespective of accuracy) is measured via the Beta statistic (β). Beta, is a statistic that provides a measurement of the extent to which one response is more probable than another and is calculated using the equation:

$$\beta = exp\{d' \ x \ C\}, \text{ where } C = \vartheta - \{d'/2\}$$

The two other measures generated by our design (and used in the calculation of Beta) are *Miss Rate* – referring to phishing emails that were identified as genuine emails and *True Rejection Rate* – genuine emails that were identified as such by the participant. We refer readers to relevant texts (e.g. [7, 32]) for further information on this method.

4.2 Sensitivity to Phishing Emails

We ran three independent t-tests on the sensitivity (*d'*) scores, comparing the control (no nudge) condition with the other experimental conditions (sender saliency nudge, receiver saliency nudge, and combined nudges). Means for d' in each condition are presented in Table 2.

Table 2: Sensitivity *d'* (higher is better) for each nudging condition (range: -4.53 - 4.53).

Nudge	N	Mean (*d'*)
None (Control)	65	0.59
Sender	64	0.98
Receiver	79	0.87
Combined	73	0.92
Total	281	0.79

These planned comparisons revealed a significant improvement in phish detection (*d'*) when sender saliency cues were employed ($t(127)=2.080, p=.020$) but no significant difference when receiver saliency cues were employed ($t(142)=1.498, p=.068$). We also found improved performance against the control when the cues were combined, i.e. when both sender and receiver salience cues were present ($t(142)=1.667, p=.049$). An additional t-test between the sender saliency and the receiver saliency cues reported no significant differences between the two ($t(141)=.598, p=.551$).

4.3 Bias

We compared the bias (β) score for each nudging condition against the control to determine whether the nudges influenced the likelihood of participants to respond "phish" or "genuine" irrespective of what was actually presented. Means for β can be seen on Table 3.

Table 3: Bias β (low = tendency to select "phish", high = tendency to select "genuine").

Nudge	N	Mean (β)
None (Control)	65	1.90
Sender	64	1.47
Receiver	79	1.87
Combined	73	1.65
Total	281	1.73

Again, planned independent t-tests were made of the experimental conditions against the control. There were no statistically significant differences when comparing the sender saliency condition ($t(127)=1.439, p=.153$), the receiver saliency condition ($t(142)=0.100, p=.920$), or the combined cues conditions ($t(136)=0.773, p=.441$) against the control. Thus, the improved detection performance for sender salience and combined conditions noted above were not associated with any change in the participants' bias in terms of a tendency to classify emails as phish or as genuine.

4.4 Decision Time

While participants' sensitivity to phishing emails and their bias were the main variables under investigation, the time taken to make each decision was considered important in the light of the drive towards productive security solutions. The time taken to decide on individual emails ranged from 3 seconds to 117 seconds, with a mode of 9 seconds. We note that only 1.3% of decisions were made in 3 seconds, evenly spread across participants. Table 4 presents the mean number of seconds required to select a response per email.

Table 4: Average time taken to make a decision on an email (seconds per email).

Nudge	N	Mean (*seconds*)
None (Control)	65	19.91
Sender	64	20.28
Receiver	79	18.18
Combined	73	18.79
Total	281	19.22

We found no significant difference in time taken to make a decision when comparing each of the experimental conditions against the control, i.e. no difference for sender salience ($t(127)=0.208, p=.836$), for receiver salience ($t(142)=0.975, p=.331$), or for the combined condition ($t(136)=0.725, p=.470$). Thus, the improvements in detection accuracy, presented above, do not incur a time penalty and should not lead to productivity losses.

4.5 Calibration of Confidence

We noted earlier the importance of well calibrated confidence in making risk decisions. In this study, we measured user confidence in each email judgment and mean confidence ratings are given in Table 5, below.

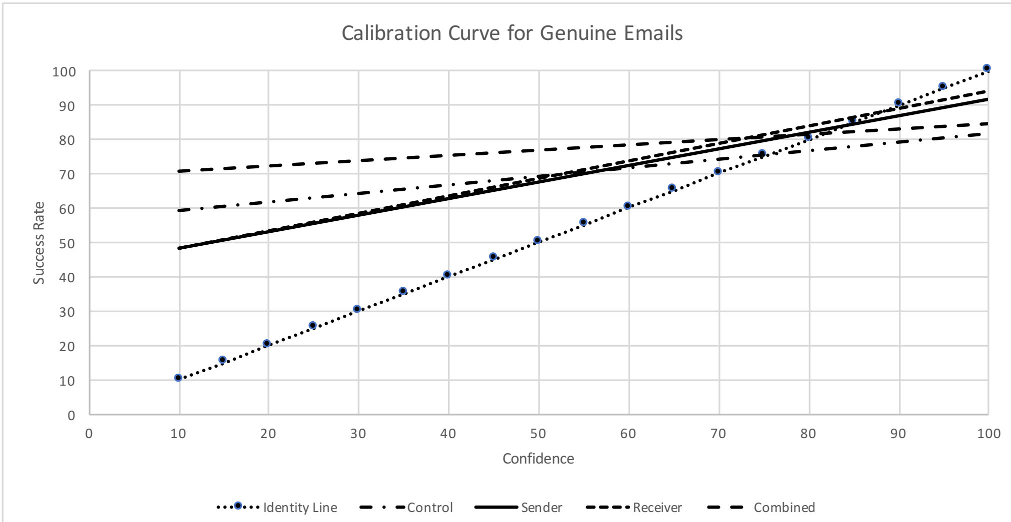


Figure 2: Calibration Curve for genuine emails. The identity line shows perfect calibration with underconfident responses plotted above and underconfident responses plotted below.

Table 5: Confidence (%) indicated by participants per choice.

Nudge	N	Mean (%)
None (Control)	65	68.24
Sender	64	69.63
Receiver	79	69.89
Combined	73	67.12
Total	281	68.73

We then constructed confidence calibration curves for both the phishing and the genuine emails. A calibration curve is a graph where subjective confidence of being correct is plotted against the actual performance (in this case percentage confidence is measured against percentage accuracy). The curves are created by computing the mean accuracy of those items where participants have given a particular confidence score. On each figure, the diagonal or *identity line* shows perfect calibration. Any data points above this line show *under-confidence* and points below the line show *over-confidence*. To take one example, a data point that shows 80% on the x-axis and 40% on the y-axis is showing that when we aggregate those emails in which the mean confidence rating is 80%, the mean accuracy rate for those same emails is only 40% (i.e. participants are over-confident). Thus good calibration would be indicated by data curves forming close to the diagonal or identity line and poor calibration would be shown by deviation from this line [30].

If we look firstly at the calibration curves for genuine emails (Figure 2) then we can see that under-confidence predominates – users are generally more accurate than they believe themselves to be. However, there appears to be a linear trend, suggesting that greater confidence is generally associated with better accuracy and there is some suggestion that the two 'nudges' of cueing sender and receiver salience can act to improve calibration of confidence.

Turning now to the calibration curves for phishing emails (Figure 3) then we can see how poorly calibrated user confidence is for

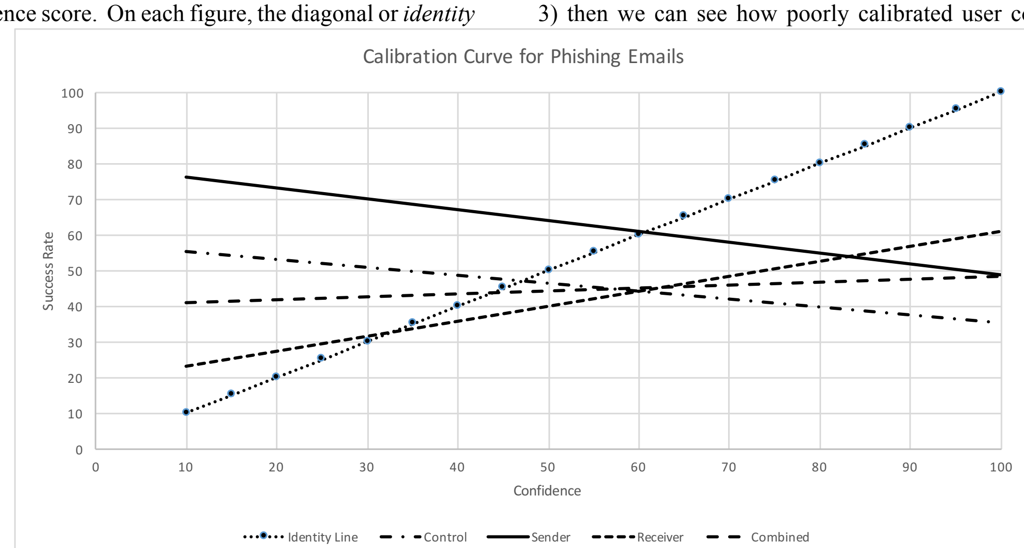


Figure 3: Calibration Curve for phishing emails. The identity line shows perfect calibration with underconfident responses plotted above and underconfident responses plotted below.

these items – with no overall sense that users are sensitive to their own ability to detect phish. The improvements in phish detection that gave rise to the significant *d'* score in the sender salience condition is reflected here in the solid line being associated with higher accuracy rates, but what is fascinating is that the cue that gives rise to improved accuracy cannot be harnessed to give users a better sense of how well they are doing in making this judgment.

4.6 Impulsivity

We used the scores on the Dickman scales to identify the top and bottom quartiles for both functional and dysfunctional impulsivity (i.e. we created four groups with approximately 60 participants per group and used the top scoring and bottom scoring groups for the analysis while discarding the middle two groups). We then conducted t-tests to compare these groups and found no significant effect of functional impulsivity (again taking *d'* as the measure of phishing sensitivity) (*t(157)=1.348, p=.179*). However, for the dysfunctional trait we found a significant difference in sensitivity to phishing emails between high and low scorers (*t(142)=2.987, p=.003*) where participants who scored high in dysfunctional impulsivity were relatively poor at detecting phish (*d'*=0.62) when compared to those with those who scored low on the trait (*d'*=1.13).

These findings beg the question of whether or not the different nudges we have designed would be particularly effective as "protective measures" for those with dysfunctional impulsivity, but here we hit an analysis problem as we have not controlled for dysfunctional impulsivity in our allocation of participants to groups and so have a variable distribution of 'dysfunctional impulsives' across cells compounded by a relatively low *n* which makes us reluctant to undertake an inferential analysis. For completeness, however, we give the sensitivity scores (*d'*) and standard deviations for each of the conditions in Table 6, below.

Table 6: Phishing sensitivity (*d'*) by condition for high and low impulsives.

Nudge	N	High dysfunctional impulsivity *d'* (s.d.)	N	Low dysfunctional impulsivity *d'* (s.d.)
None (Control)	14	0.15 (0.97)	12	0.67 (1.08)
Sender	15	0.98 (1.17)	11	1.12 (0.63)
Receiver	16	0.58 (1.25)	15	1.60 (1.07)
Combined	19	0.79 (1.21)	21	1.19 (1.44)

Finally, we found a significant difference between those scoring high vs. low on dysfunctional impulsivity on the time taken to make decisions, using a non-parametric Mann Whitney U test due to the non-normal distribution of data (*U=1803.5, p<0.01*). High dysfunctional impulsives made faster decisions on average (mean = 15.90s) than low dysfunctional impulsives (mean = 18.27s).

5. Discussion

The purpose of this study was to evaluate whether we can use the social context of sending and receiving emails to improve participants' ability to detect phish. We highlighted information about either sender (name, email address, and the time the email was sent) or receiver (number of people in the organisation who received the email) as two means of *nudging* people to think more carefully about the communicative context of the email. These two nudges individually and in combination were tested against a control where users were simply shown the email in its original, non-altered format.

We found that improving sender saliency led to better phish detection when compared with a control and that sender and receiver nudges used in combination also improved performance, although there was no real sense of any added value from the receiver nudge. The improvements were not associated with any overall bias in terms of participants' inclination to decide "phish" or "genuine". We also found that participants were under-confident in their decisions when presented with genuine emails, but were over-confident when presented with phishing emails. Finally, we found that participants who scored high on the trait of dysfunctional impulsivity [13] were less accurate in identifying phishing emails and made faster decisions than those scoring low for the trait. These results are discussed in more detail below.

5.1 Performance with Nudges

The sender saliency nudge presented alone and in combination with a receiver saliency nudge improved phish detection over the control condition. In other words, the simple act of highlighting fields that are already present in an email – sender's name, email address and time sent – was an effective means of improving user security – a finding that is consistent with other work that suggests persuading users to attend to such information can help users with phish identification [16]. Users already rate these features as important for identifying phishing emails, with 95% of lay participants reporting that they use the "from" field to pick out discrepancies between email and sender name [15]. However, it seems that this knowledge is not being applied in practice – even under those circumstances where participants had been instructed to look for phishing emails. We should also note that the email address field is by default hidden in several popular email clients. For example, on Gmail's web interface a user is required to hover over the sender's name in order to bring up the email address (after a few seconds' delay), and in Microsoft Outlook the user has to perform a number of steps in order to be able to see the origin email address. These practices are unlikely to help users in spotting discrepancies in emails and should be avoided.

Our results show that participants using the receiver saliency nudge (i.e. indicating how many other people were in receipt of the email) did not perform significantly better than participants viewing the email without nudges (control). It is possible that the wording used for the receiver saliency condition – highlighting the percentage of "colleagues" who also received the email – was not descriptive enough for participants, and a more detailed approach similar to that employed by Wang et al. [47] where specific individuals are named may work better. However, the privacy implications of such an implementation in an organisation should first be considered.

We did not find any associated effect on bias (i.e. participants were no more likely to select "phish" overall when nudged, irrespective of whether the email was or was not genuine). This is important, as nudges that simply make people more or less conservative overall (without improving sensitivity) could have unfortunate consequences, leading to either the rejection of genuine emails or the acceptance of fraudulent emails.

5.2 Confidence in Phish Detection

Parsons et al. [34] have shown that participants are more accurate at identifying phishing emails when they know they are taking part in a phishing experiment. However, our participants were rather poor at phish detection, and more worrying, were not well calibrated in terms of the confidence they placed in their own judgements, further supporting previous work emphasising the importance of self-confidence when identifying phishing emails [7]. In other words, there was a discrepancy between subjective confidence and objective performance when classifying emails and this discrepancy seemed particularly problematic for phishing emails, where participants were generally poor at detecting phish (i.e. showed lower accuracy levels) but were overconfident that they made the right decision. On the other hand, for genuine emails, participants showed better calibration in confidence scores, although showed an overall pattern of under-confidence. This finding is interesting and is probably worth pursuing further. It is conceivable that users employ different cues for the detection of genuine emails than they do for the detection of phish but we would need to explore this issue in future studies. For the moment, we might note that good calibration of confidence essentially depends on both the amount and the strength of the evidence available in supporting the choice [27]. For phish decisions, users have relatively poor sources of evidence available and this is likely to be compounded by their inability to assess the quality of that evidence.

5.3 Impulsivity and Phish Detection

Previous work has suggested that impulsivity may play a role in phish detection. For instance, less impulsive people have been found to manage email better (i.e. spot phishing more efficiently) than those scoring high for the trait, based on the Cognitive Reflection Task [35]. It should be noted, however, that this result pertained to participants who were aware that they were taking part in a phishing task (i.e. were vigilant). Kumaraguru et al. [28] also found a trend where participants with lower Cognitive Reflective Task scores (i.e. with higher impulsivity) were more likely to click on phishing emails, although in this case the trend was not statistically significant.

Our results are consistent with these findings, but here we have used Dickman's distinction between functional and dysfunctional impulsivity, finding that only the dysfunctional scale is associated with poor phish detection. What is encouraging, is that our sender saliency intervention would appear to be effective even for those with low impulsivity (Table 5) however we have been reluctant to conduct any inferential statistics on these data as the power would be rather low, given the relatively small cell sizes and of course we have not systematically controlled for levels of impulsivity across the intervention conditions.

5.4 The Use of Signal Detection Theory in Phishing Research

In the past, phishing research and email classification in general typically analysed results using separate measures for success rate and false positives (e.g. [40]) or simply an accuracy measurement (e.g. [14, 19, 21, 28, 35]). This results in a simple ratio that indicates how comfortable users are identifying phish but neglects false alarms (i.e. incorrectly classifying a genuine email as a phish). Yet false alarms are becoming a concern for organisations as they are associated with productivity and/or business losses that arise when staff ignore legitimate emails. Additionally, simple measures of success that ignore decision bias are also problematic as changing the tendency to classify emails as phish or genuine irrespective of their actual legitimacy is not the target outcome.

Signal detection theory accounts for both false positives and response bias with the two main measures of sensitivity (d') and bias (β). We have shown in this paper how applying this analysis method teases out intricate performance measures that may be missed when using other methods. We are aware of two other papers that have recently utilised signal detection theory in phish detection [7, 34]. Canfield et al. [7] found that participants were accurate in determining the correct action for phishing emails (deleting or marking as "spam"), but that their sensitivity to phishing emails was poor. Parsons et al. [34] found that participants aware of their participation in phishing experiments were more sensitive to the phishing emails.

We are pleased that this measure is being adopted in phishing research, given how important the separation of sensitivity and bias are for realistic interventions in phishing.

5.5 Limitations

Although we were able to obtain a number of interesting insights from the study, there are two main limitations that we should discuss that may have affected the performance of participants.

Firstly, the messages used for both phishing and genuine emails were not actually received by the participants, thus it is unclear how familiar they were with each email. For example, it is possible that some participants may be familiar with the receipt of Amazon emails, direct from the retailer. They could then have used this knowledge to help them pick up subtle cues to aid their decision making. Whereas other participants may be unfamiliar with Amazon and as a result be at a disadvantage when judging the veracity of emails. This is a common pitfall with lab-based phishing experiments and can be addressed by running "in the wild" studies, although these introduce other limitations.

Secondly, participants were unable to interact directly with the email messages or carry out any additional checks (e.g. search for the company online). We addressed this issue by always having visible cues to allow informed decisions (see Section 3.3 for details).

Finally, participants were told from the beginning that they were taking part in a phishing experiment. These instructions will have primed them to scrutinise each email more closely than perhaps they would do otherwise [34]. However, given that all participants in all conditions were subjected to these instructions then this should not have affected our main findings – i.e. the effects should be the same for all. We note that, overall, our participants may have shown better sensitivity to phish than those who receive fraudulent emails "in the wild". Unfortunately, we cannot compare our findings with any normative data as sensitivity estimates are not available elsewhere.

6. Conclusions

In this paper, we evaluated two nudges with the aim of improving users' phishing detection on email clients. We found that users were more successful identifying phishing emails when their attention was drawn to the sender's details (name and originating email address) and the time received when compared with the control condition. This is problematic, given the recent design trend on popular email clients to hide important sender information (i.e. the full originating email address) by default, thus potentially hindering users' efforts when evaluating emails in their everyday lives. We also found strong evidence that individuals scoring high for dysfunctional impulsivity were at a disadvantage when identifying phishing emails and set this finding against previous published work which has been inconclusive about the effect of impulsive behaviour on phishing identification (e.g. [28, 35]). We noted an interesting finding in relationship to users' overconfidence when making decisions in respect of phishing emails (and underconfidence in respect of genuine emails) and we would encourage further research in this area. Finally, we would recommend the adoption of Signal Detection Theory for phishing research, in particular due to the response bias measure that allows further scrutiny of potential interventions.

7. ACKNOWLEDGMENTS

The work presented in this paper was funded through the Choice Architecture for Information Security (ChAIse) project (EP/K006568/1) from Engineering and Physical Sciences Research Council (EPSRC), UK, and Government Communications Headquarters (GCHQ), UK, as a part of the Research Institute in Science of Cyber Security.

8. REFERENCES

[1] Almuhimedi, H., Schaub, F., Sadeh, N., Adjerid, I., Acquisti, A., Gluck, J., Cranor, L.F. and Agarwal, Y. 2015. Your Location has been Shared 5,398 Times! *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15* (New York, New York, USA, 2015), 787–796.

[2] Beautement, A., Sasse, M.A. and Wonham, M. 2008. The compliance budget. *Proceedings of the 2008 workshop on New security paradigms - NSPW '08* (New York, New York, USA, Aug. 2008), 47.

[3] Bergholz, A., De Beer, J., Glahn, S., Moens, M.-F., Paaß, G. and Strobel, S. 2010. New filtering approaches for phishing email. *Journal of Computer Security*. 18, 1 (Jan. 2010), 7–35.

[4] Bravo-Lillo, C., Komanduri, S., Cranor, L.F., Reeder, R.W., Sleeper, M., Downs, J. and Schechter, S. 2013. Your attention please. *Proceedings of the Ninth Symposium on Usable Privacy and Security - SOUPS '13* (New York, New York, USA, 2013), 1.

[5] Brustoloni, J.C. and Villamar'\in-Salomón, R. 2007. Improving security decisions with polymorphic and audited dialogs. *Proceedings of the 3rd symposium on Usable privacy and security*. (2007), 76–85.

[6] Bursztein, E., Margolis, D., Archer, A., Pitsillidis, A. and Savage, S. 2014. Handcrafted Fraud and Extortion : Manual Account Hijacking in the Wild. *In Proceedings of the 2014 Conference on Internet Measurement Conference* (2014), 347–358.

[7] Canfield, C.I., Fischhoff, B. and Davis, A. 2016. Quantifying Phishing Susceptibility for Detection and Behavior Decisions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*. 58, 8 (2016), 1158–1172.

[8] Caputo, D.D., Pfleeger, S.L., Freeman, J.D. and Johnson, M.E. 2014. Going Spear Phishing: Exploring Embedded Training and Awareness. *IEEE Security & Privacy*. 12, 1 (Jan. 2014), 28–38.

[9] Chandrasekaran, M., Narayanan, K. and Upadhyaya, S. 2006. Phishing E-mail Detection Based on Structural Properties. *NYS Cyber Security Conference* (2006), 1–7.

[10] Chen, J., Gates, C.S., Li, N. and Proctor, R.W. 2015. Influence of Risk/Safety Information Framing on Android App-Installation Decisions. *Journal of Cognitive Engineering and Decision Making*. 9, 2 (Jun. 2015), 149–168.

[11] Choe, E.K., Jung, J., Lee, B. and Fisher, K. 2013. Nudging People Away from Privacy-Invasive Mobile Apps through Visual Framing. Springer Berlin Heidelberg. 74–91.

[12] Chou, N., Ledesma, R., Teraguchi, Y. and Mitchell, J.C. 2004. Client-side defense against web-based identity theft. *Most*. (2004), 1–16.

[13] Dickman, S.J. 1990. Functional and dysfunctional impulsivity: personality and cognitive correlates. *Journal of personality and social psychology*. 58, 1 (Jan. 1990), 95–102.

[14] Dodge, R.C., Carver, C. and Ferguson, A.J. 2007. Phishing for user security awareness. *Computers & Security*. 26, 1 (2007), 73–80.

[15] Downs, J.S., Holbrook, M.B. and Cranor, L.F. 2006. Decision strategies and susceptibility to phishing. *Proceedings of the second symposium on Usable privacy and security - SOUPS '06* (New York, New York, USA, 2006), 79.

[16] Drake, C.E., Oliver, J.J. and Koontz, E.J. 2004. Anatomy of a Phishing Email. *Proceedings of the First Conference on E-mail and Anti-Spam (CEAS)* (2004), 1–8.

[17] Egelman, S., Cranor, L.F. and Hong, J. 2008. You've been warned. *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08* (New York, New York, USA, 2008), 1065.

[18] Even security experts fail to spot phishing emails, finds report: 2015. .

[19] Ferguson, A.J. 2005. Fostering e-mail security awareness: The West Point carronade. *Educase Quarterly*. 28, 1 (2005), 54–57.

[20] Ferreira, A., Coventry, L. and Lenzini, G. 2015. Principles of persuasion in social engineering and their use in phishing. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2015), 36–47.

[21] Furnell, S. 2007. Phishing: can we spot the signs? *Computer Fraud & Security*. 2007, 3 (2007), 10–15.

[22] Grazioli, S. 2004. Where Did They Go Wrong? An Analysis of the Failure of Knowledgeable Internet Consumers to Detect Deception Over the Internet. *Group Decision and Negotiation*. 13, 2 (Mar. 2004), 149–172.

[23] Halevi, T., Lewis, J. and Memon, N. 2013. A pilot study of cyber security and privacy related behavior and personality traits. *Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion* (New York, New York, USA, 2013), 737–744.

[24] Herley, C. 2009. 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 (NSPW '09) (2009), 133–144.

[25] Hong, K.W., Kelley, C.M., Tembe, R., Murphy-Hill, E. and Mayhorn, C.B. 2013. Keeping Up With The Joneses: Assessing Phishing Susceptibility in an Email Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 57, 1 (Sep. 2013), 1012–1016.

[26] Jagatic, T.N., Johnson, N.A., Jakobsson, M. and Menczer, F. 2007. Social phishing. *Communications of the ACM*. 50, (2007), 94–100.

[27] Koriat, A., Lichtenstein, S. and Fischhoff, B. 1980. Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory*. 6, 2 (1980), 107–118.

[28] Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L.F., Hong, J. and Nunge, E. 2007. Protecting people from phishing: the design and evaluation of an embedded training email system. *Proceedings of ACM CHI 2007 Conference on Human Factors in Computing Systems*. 1, (2007), 905–914.

[29] Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L.F. and Hong, J. 2010. Teaching Johnny not to fall for phish. *ACM Transactions on Internet Technology*. 10, 2 (May 2010), 1–31.

[30] Lichtenstein, S. and Fischhoff, B. 1977. Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*. 20, 2 (Dec. 1977), 159–183.

[31] McAfee Labs Threat Report August 2014: 2014. *http://www.mcafee.com/uk/security-awareness/articles/mcafee-labs-threats-report-q2-2014.aspx*.

[32] McNicol, D. 2005. *A Primer of Signal Detection Theory*. Taylor and Francis.

[33] Modic, D. and Lea, S.E.G. 2011. How Neurotic are Scam Victims, Really? The Big Five and Internet Scams. *In Proceedings of the Conference of the International Confederation for the Advancement of Behavioral Economics and Economic Psychology* (2011).

[34] Parsons, K., McCormac, A., Pattinson, M., Butavicius, M. and Jerram, C. 2014. The design of phishing studies: Challenges for researchers. *Computers and Security*. (2014).

[35] Pattinson, M., Jerram, C., Parsons, K., McCormac, A. and Butavicius, M. 2012. Why do some people manage phishing e-mails better than others? *Information Management & Computer Security*. 20, (2012), 18–28.

[36] Riek, M., Bohme, R. and Moore, T. 2016. Measuring the Influence of Perceived Cybercrime Risk on Online Service Avoidance. *IEEE Transactions on Dependable and Secure Computing*. 13, 2 (Mar. 2016), 261–273.

[37] RSA Online Fraud Report 2014: 2014. *https://www.rsa.com/de-de/perspectives/resources/2014-cybercrime-roundup*.

[38] Safe browsing - transparency report - Google: 2013. *https://www.google.com/transparencyreport/safebrowsing/*.

[39] Sasse, M.A., Brostoff, S. and Weirich, D. 2001. Transforming the "weakest link" — a human/computer interaction approach to usable and effective security. *BT Technology Journal*. 19, 3 (2001), 122–131.

[40] Sheng, S., Holbrook, M., Kumaraguru, P., Cranor, L.F. and Downs, J. 2010. Who falls for phish? A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions. *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10* (2010), 373–382.

[41] Sheng, S., Magnien, B., Kumaraguru, P., Acquisti, A., Cranor, L.F., Hong, J. and Nunge, E. 2007. Anti-Phishing Phil: the design and evaluation of a game that teaches people not to fall for phish. *Proceedings of the 3rd Symposium on Usable Privacy and Security - SOUPS '07* (New York, New York, USA, 2007), 88.

[42] Symantec Corporation 2014. Internet Security Threat Report. 19, April (2014), 97.

[43] Thaler, R.H. and Sunstein, C.R. 2009. *Nudge: Improving decisions about health, wealth, and happiness*. Yale.

[44] "This message may not have been sent by..." warning: 2016. *https://support.google.com/mail/troubleshooter/2411000?hl=en*.

[45] Thunderbird's Scam Detection: 2016. *https://support.mozilla.org/en-US/kb/thunderbirds-scam-detection#w_thunderbirds-automatic-scam-filtering*.

[46] Wang, P.A. 2013. Assessment of Cybersecurity Knowledge and Behavior : An Anti-phishing Scenario. *ICIMP 2013: The Eighth International Conference on Internet Monitoring and Protection*. c (2013), 1–7.

[47] Wang, Y., Leon, P.G., Acquisti, A., Cranor, L.F., Forget, A., Sadeh, N., Wang, Y., Leon, P.G., Acquisti, A., Cranor, L.F., Forget, A. and Sadeh, N. 2014. A field trial of privacy nudges for facebook. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14* (New York, New York, USA, 2014), 2367–2376.

[48] Wu, M., Miller, R.C. and Garfinkel, S.L. 2006. Do security toolbars actually prevent phishing attacks? *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06* (2006), 601.

[49] Wu, S.-Y. 2009. *PhishDuck : Capturing User Intention in an Email Client to Combat Phishing*. Carnegie Mellon University.

[50] Zhang, Y., Egelman, S., Cranor, L. and Hong, J. 2006. Phinding Phish: Evaluating Anti-Phishing Tools. (2006).

[51] Zhang, Y., Hong, J.I. and Cranor, L.F. 2007. Cantina: a content-based approach to detecting phishing web sites. *Proceedings of the 16th international conference on World Wide Web - WWW '07*. (2007), 639.

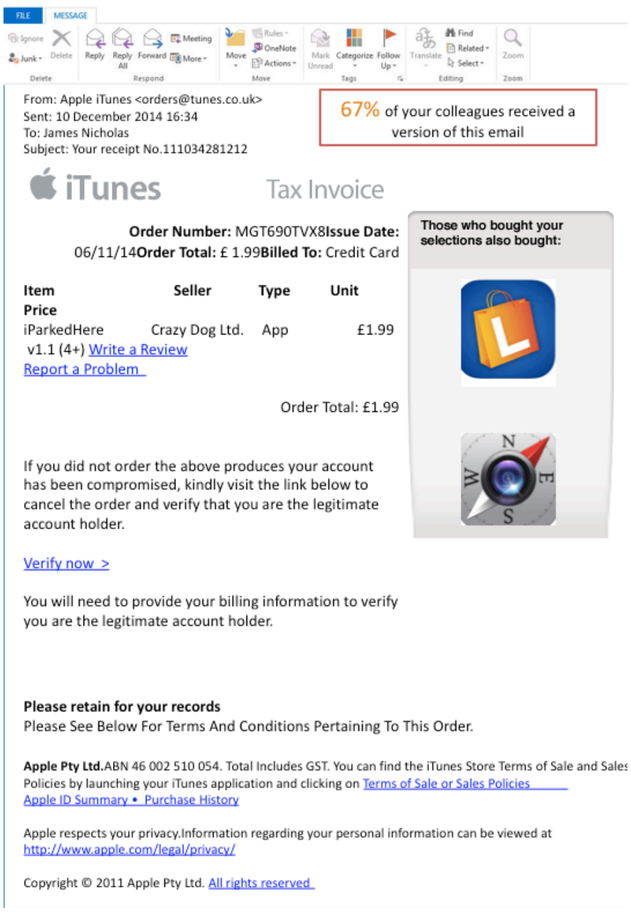
A. APPENDIX

A.1 Email Screenshots

The sender saliency nudge emails used for the experiment are presented below. The same set of emails were used for the control condition (without any mark-up), the sender saliency condition (without the percentage mark-up and with added highlights over the key metadata – see Figure 1) and the combined condition (with added highlights over the key metadata – see Figure 1).

Phishing Emails

Below are the six phishing emails, collected from existing messages that were found problematic by our university.


From: Apple iTunes <orders@tunes.co.uk>
Sent: 10 December 2014 16:34
To: James Nicholas
Subject: Your receipt No.111034281212

67% of your colleagues received a version of this email

iTunes Tax Invoice

Order Number: MGT690TVX8Issue Date: 06/11/14Order Total: £ 1.99Billed To: Credit Card

Those who bought your selections also bought:

Item	Seller	Type	Unit Price
iParkedHere v1.1 (4+) Write a Review Report a Problem	Crazy Dog Ltd.	App	£1.99

Order Total: £1.99

If you did not order the above produces your account has been compromised, kindly visit the link below to cancel the order and verify that you are the legitimate account holder.

Verify now >

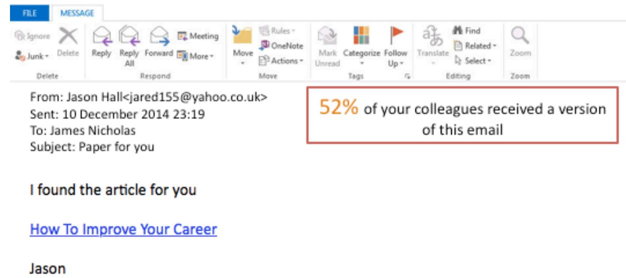
You will need to provide your billing information to verify you are the legitimate account holder.

Please retain for your records
Please See Below For Terms And Conditions Pertaining To This Order.

Apple Pty Ltd.ABN 46 002 510 054. Total Includes GST. You can find the iTunes Store Terms of Sale and Sales Policies by launching your iTunes application and clicking on Terms of Sale or Sales Policies
Apple ID Summary • Purchase History

Apple respects your privacy.Information regarding your personal information can be viewed at http://www.apple.com/legal/privacy/

Copyright © 2011 Apple Pty Ltd. All rights reserved

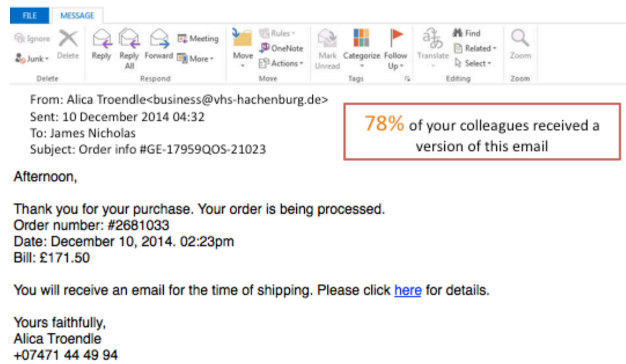

From: Jason Hall<jared155@yahoo.co.uk>
Sent: 10 December 2014 23:19
To: James Nicholas
Subject: Paper for you

52% of your colleagues received a version of this email

I found the article for you

How To Improve Your Career

Jason


From: Alica Troendle<business@vhs-hachenburg.de>
Sent: 10 December 2014 04:32
To: James Nicholas
Subject: Order info #GE-17959QOS-21023

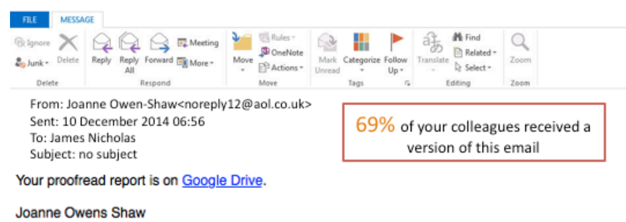
78% of your colleagues received a version of this email

Afternoon,

Thank you for your purchase. Your order is being processed.
Order number: #2681033
Date: December 10, 2014. 02:23pm
Bill: £171.50

You will receive an email for the time of shipping. Please click here for details.

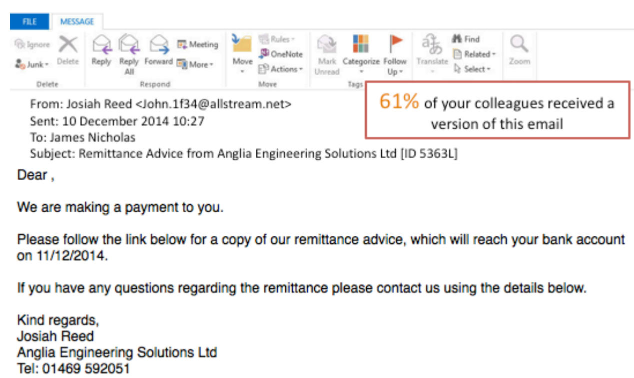
Yours faithfully,
Alica Troendle
+07471 44 49 94


From: Joanne Owen-Shaw<noreply12@aol.co.uk>
Sent: 10 December 2014 06:56
To: James Nicholas
Subject: no subject

69% of your colleagues received a version of this email

Your proofread report is on Google Drive.

Joanne Owens Shaw


From: Josiah Reed <John.1f34@allstream.net>
Sent: 10 December 2014 10:27
To: James Nicholas
Subject: Remittance Advice from Anglia Engineering Solutions Ltd [ID 5363L]

61% of your colleagues received a version of this email

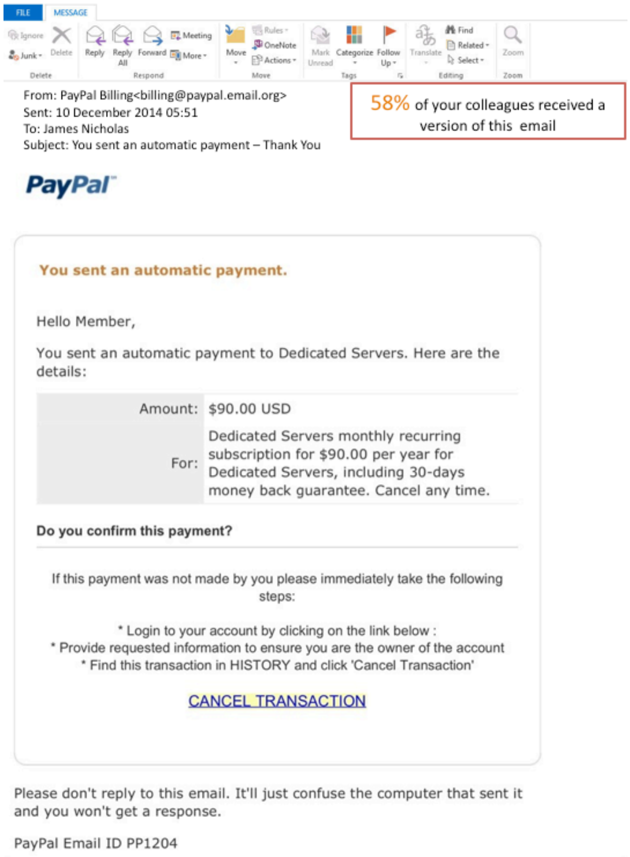
Dear ,

We are making a payment to you.

Please follow the link below for a copy of our remittance advice, which will reach your bank account on 11/12/2014.

If you have any questions regarding the remittance please contact us using the details below.

Kind regards,
Josiah Reed
Anglia Engineering Solutions Ltd
Tel: 01469 592051



From: PayPal Billing<billing@paypal.email.org>
Sent: 10 December 2014 05:51
To: James Nicholas
Subject: You sent an automatic payment – Thank You

58% of your colleagues received a version of this email

PayPal

You sent an automatic payment.

Hello Member,

You sent an automatic payment to Dedicated Servers. Here are the details:

Amount:	$90.00 USD
For:	Dedicated Servers monthly recurring subscription for $90.00 per year for Dedicated Servers, including 30-days money back guarantee. Cancel any time.

Do you confirm this payment?

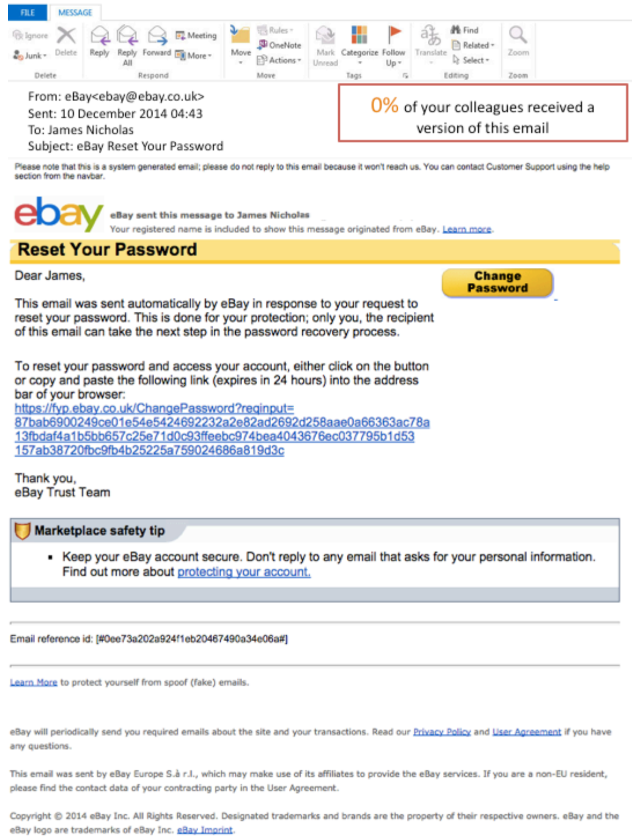
If this payment was not made by you please immediately take the following steps:

* Login to your account by clicking on the link below :
* Provide requested information to ensure you are the owner of the account
* Find this transaction in HISTORY and click 'Cancel Transaction'

CANCEL TRANSACTION

Please don't reply to this email. It'll just confuse the computer that sent it and you won't get a response.

PayPal Email ID PP1204



From: eBay<ebay@ebay.co.uk>
Sent: 10 December 2014 04:43
To: James Nicholas
Subject: eBay Reset Your Password

0% of your colleagues received a version of this email

Please note that this is a system generated email; please do not reply to this email because it won't reach us. You can contact Customer Support using the help section from the navbar.

ebay **eBay sent this message to James Nicholas**
Your registered name is included to show this message originated from eBay. Learn more.

Reset Your Password

Dear James,

Change Password

This email was sent automatically by eBay in response to your request to reset your password. This is done for your protection; only you, the recipient of this email can take the next step in the password recovery process.

To reset your password and access your account, either click on the button or copy and paste the following link (expires in 24 hours) into the address bar of your browser:
https://fyp.ebay.co.uk/ChangePassword?reqinput=
87bab6900249ce01e54e5424692232a2e82ad2692d258aae0a66363ac78a
13fbdaf4a1b5bb657c25e71d0c93ffeebc974bea4043676ec037795b1d53
157ab38720fbc9fb4b25225a759024686a819d3c

Thank you,
eBay Trust Team

Marketplace safety tip

- Keep your eBay account secure. Don't reply to any email that asks for your personal information. Find out more about protecting your account.

Email reference id: [#0ee73a202a924f1eb20467490a34e06a#]

Learn More to protect yourself from spoof (fake) emails.

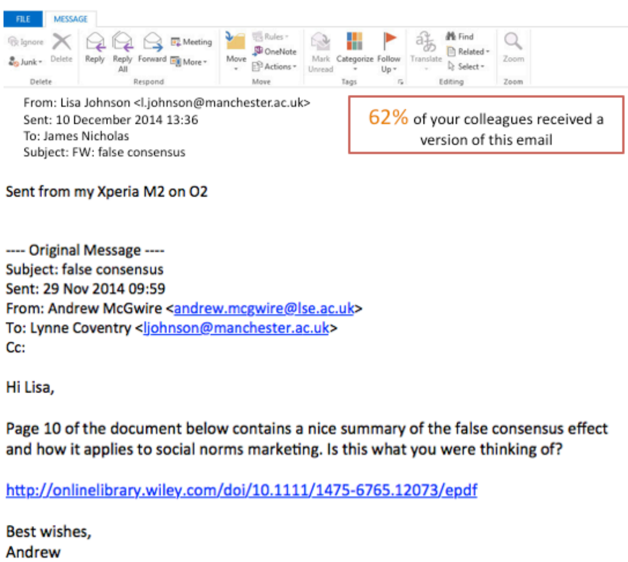
eBay will periodically send you required emails about the site and your transactions. Read our Privacy Policy and User Agreement if you have any questions.

This email was sent by eBay Europe S.à r.l., which may make use of its affiliates to provide the eBay services. If you are a non-EU resident, please find the contact data of your contracting party in the User Agreement.

Copyright © 2014 eBay Inc. All Rights Reserved. Designated trademarks and brands are the property of their respective owners. eBay and the eBay logo are trademarks of eBay Inc. eBay Imprint.

Genuine Emails

Below are the twelve genuine emails, collected by the authors.



From: Lisa Johnson <l.johnson@manchester.ac.uk>
Sent: 10 December 2014 13:36
To: James Nicholas
Subject: FW: false consensus

62% of your colleagues received a version of this email

Sent from my Xperia M2 on O2

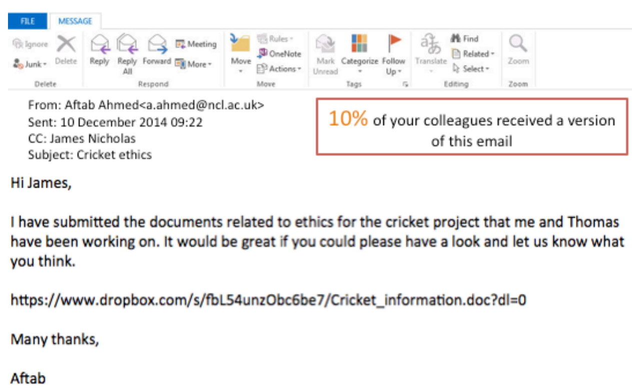
---- Original Message ----
Subject: false consensus
Sent: 29 Nov 2014 09:59
From: Andrew McGwire <andrew.mcgwire@lse.ac.uk>
To: Lynne Coventry <ljohnson@manchester.ac.uk>
Cc:

Hi Lisa,

Page 10 of the document below contains a nice summary of the false consensus effect and how it applies to social norms marketing. Is this what you were thinking of?

http://onlinelibrary.wiley.com/doi/10.1111/1475-6765.12073/epdf

Best wishes,
Andrew



From: Aftab Ahmed<a.ahmed@ncl.ac.uk>
Sent: 10 December 2014 09:22
CC: James Nicholas
Subject: Cricket ethics

10% of your colleagues received a version of this email

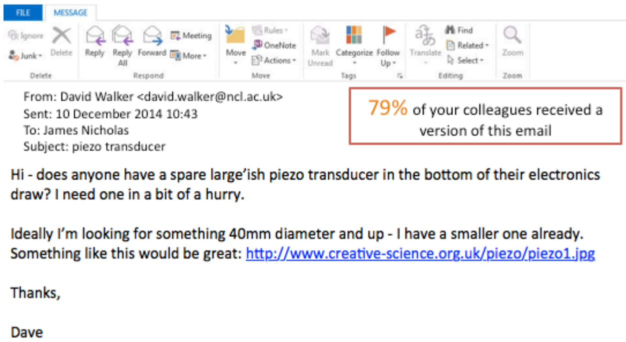
Hi James,

I have submitted the documents related to ethics for the cricket project that me and Thomas have been working on. It would be great if you could please have a look and let us know what you think.

https://www.dropbox.com/s/fbL54unzObc6be7/Cricket_information.doc?dl=0

Many thanks,

Aftab



From: David Walker <david.walker@ncl.ac.uk>
Sent: 10 December 2014 10:43
To: James Nicholas
Subject: piezo transducer

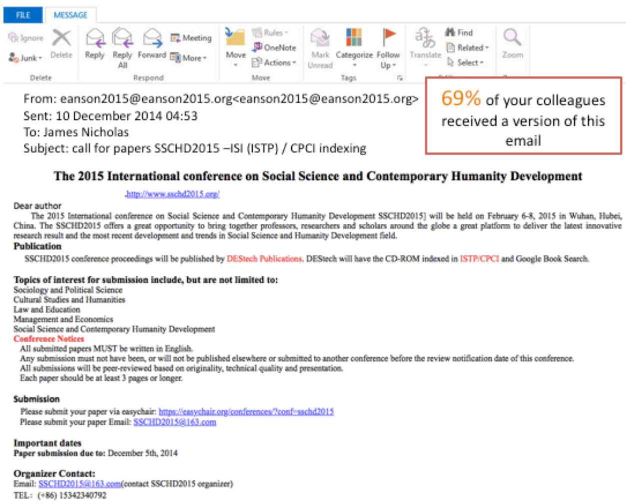
79% of your colleagues received a version of this email

Hi - does anyone have a spare large'ish piezo transducer in the bottom of their electronics draw? I need one in a bit of a hurry.

Ideally I'm looking for something 40mm diameter and up - I have a smaller one already. Something like this would be great: http://www.creative-science.org.uk/piezo/piezo1.jpg

Thanks,

Dave



From: eanson2015@eanson2015.org<eanson2015@eanson2015.org>
Sent: 10 December 2014 04:53
To: James Nicholas
Subject: call for papers SSCHD2015 –ISI (ISTP) / CPCI indexing

69% of your colleagues received a version of this email

The 2015 International conference on Social Science and Contemporary Humanity Development

http://www.sschd2015.org/

Dear author

The 2015 International conference on Social Science and Contemporary Humanity Development SSCHD2015] will be held on February 6-8, 2015 in Wuhan, Hubei, China. The SSCHD2015 offers a great opportunity to bring together professors, researchers and scholars around the globe a great platform to deliver the latest innovative research result and the most recent development and trends in Social Science and Humanity Development field.

Publication

SSCHD2015 conference proceedings will be published by DEStech Publications. DEStech will have the CD-ROM indexed in ISTP/CPCI and Google Book Search.

Topics of interest for submission include, but are not limited to:

Sociology and Political Science
Cultural Studies and Humanities
Law and Education
Management and Economics
Social Science and Contemporary Humanity Development

Conference Notices

All submitted papers MUST be written in English.
Any submission must not have been, or will not be published elsewhere or submitted to another conference before the review notification date of this conference.
All submissions will be peer-reviewed based on originality, technical quality and presentation.
Each paper should be at least 3 pages or longer.

Submission

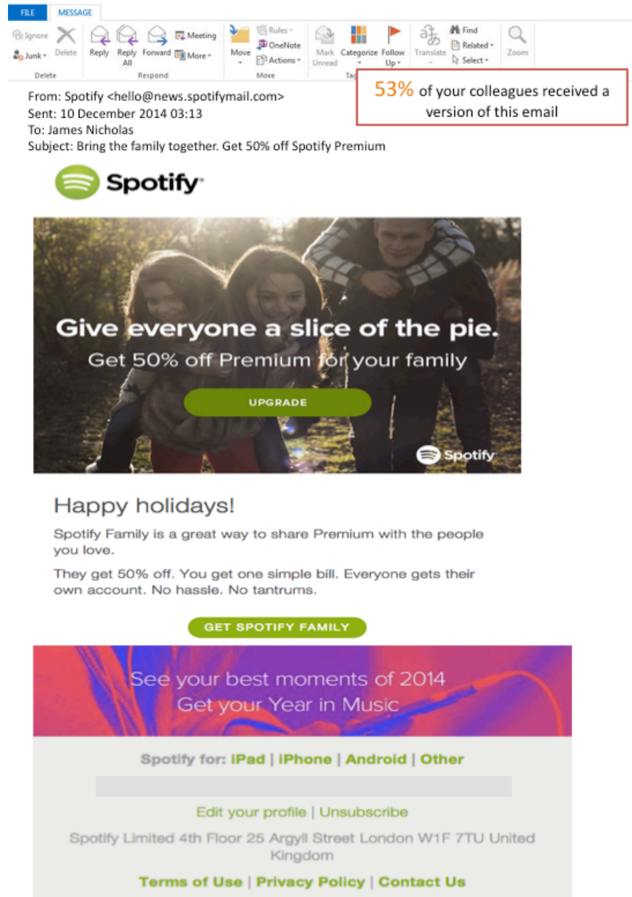
Please submit your paper via easychair: https://easychair.org/conferences/?conf=sschd2015
Please submit your paper Email: SSCHD2015@163.com

Important dates

Paper submission due to: December 5th, 2014

Organizer Contact:

Email: SSCHD2015@163.com(contact SSCHD2015 organizer)
TEL： (+86) 15342340792



From: Spotify <hello@news.spotifymail.com>
Sent: 10 December 2014 03:13
To: James Nicholas
Subject: Bring the family together. Get 50% off Spotify Premium

53% of your colleagues received a version of this email

Spotify

Give everyone a slice of the pie.
Get 50% off Premium for your family
UPGRADE
Spotify

Happy holidays!

Spotify Family is a great way to share Premium with the people you love.

They get 50% off. You get one simple bill. Everyone gets their own account. No hassle. No tantrums.

GET SPOTIFY FAMILY

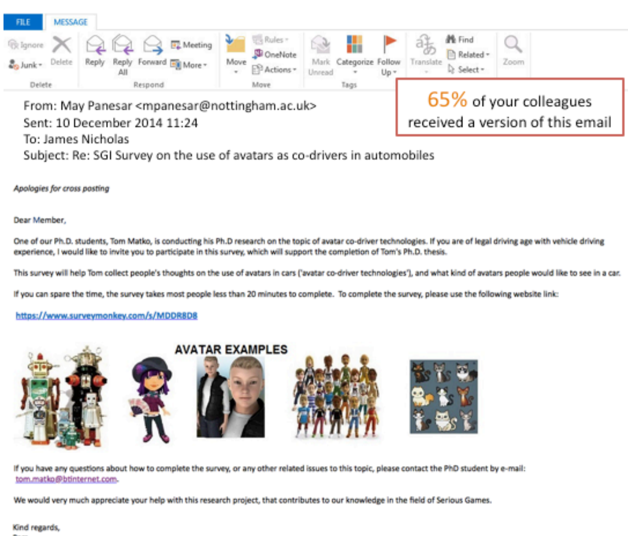
See your best moments of 2014
Get your Year in Music

Spotify for: iPad | iPhone | Android | Other

Edit your profile | Unsubscribe

Spotify Limited 4th Floor 25 Argyll Street London W1F 7TU United Kingdom

Terms of Use | Privacy Policy | Contact Us



From: May Panesar <mpanesar@nottingham.ac.uk>
Sent: 10 December 2014 11:24
To: James Nicholas
Subject: Re: SGI Survey on the use of avatars as co-drivers in automobiles

65% of your colleagues received a version of this email

Apologies for cross posting

Dear Member,

One of our Ph.D. students, Tom Matko, is conducting his Ph.D research on the topic of avatar co-driver technologies. If you are of legal driving age with vehicle driving experience, I would like to invite you to participate in this survey, which will support the completion of Tom's Ph.D. thesis.

This survey will help Tom collect people's thoughts on the use of avatars in cars ('avatar co-driver technologies'), and what kind of avatars people would like to see in a car.

If you can spare the time, the survey takes most people less than 20 minutes to complete. To complete the survey, please use the following website link:

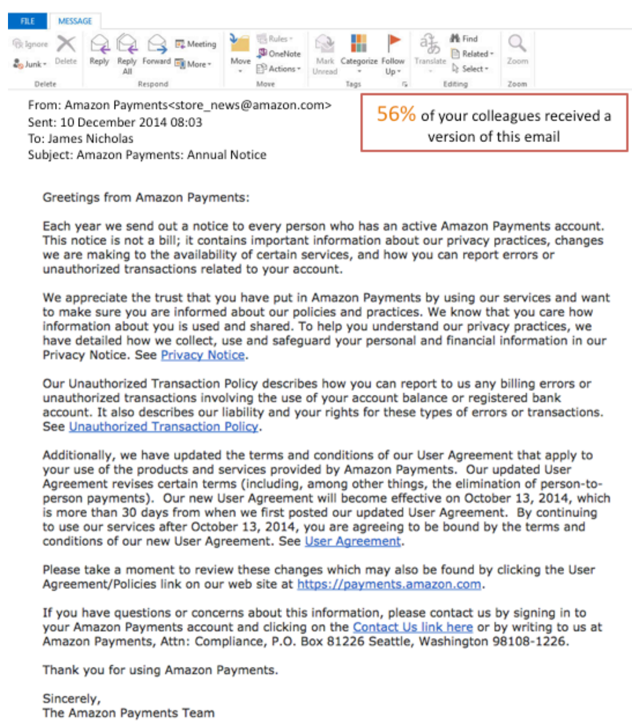
https://www.surveymonkey.com/s/MDDR8D8

AVATAR EXAMPLES

If you have any questions about how to complete the survey, or any other related issues to this topic, please contact the PhD student by e-mail: tom.matko@btinternet.com.

We would very much appreciate your help with this research project, that contributes to our knowledge in the field of Serious Games.

Kind regards,
Pam



From: Amazon Payments<store_news@amazon.com>
Sent: 10 December 2014 08:03
To: James Nicholas
Subject: Amazon Payments: Annual Notice

56% of your colleagues received a version of this email

Greetings from Amazon Payments:

Each year we send out a notice to every person who has an active Amazon Payments account. This notice is not a bill; it contains important information about our privacy practices, changes we are making to the availability of certain services, and how you can report errors or unauthorized transactions related to your account.

We appreciate the trust that you have put in Amazon Payments by using our services and want to make sure you are informed about our policies and practices. We know that you care how information about you is used and shared. To help you understand our privacy practices, we have detailed how we collect, use and safeguard your personal and financial information in our Privacy Notice. See Privacy Notice.

Our Unauthorized Transaction Policy describes how you can report to us any billing errors or unauthorized transactions involving the use of your account balance or registered bank account. It also describes our liability and your rights for these types of errors or transactions. See Unauthorized Transaction Policy.

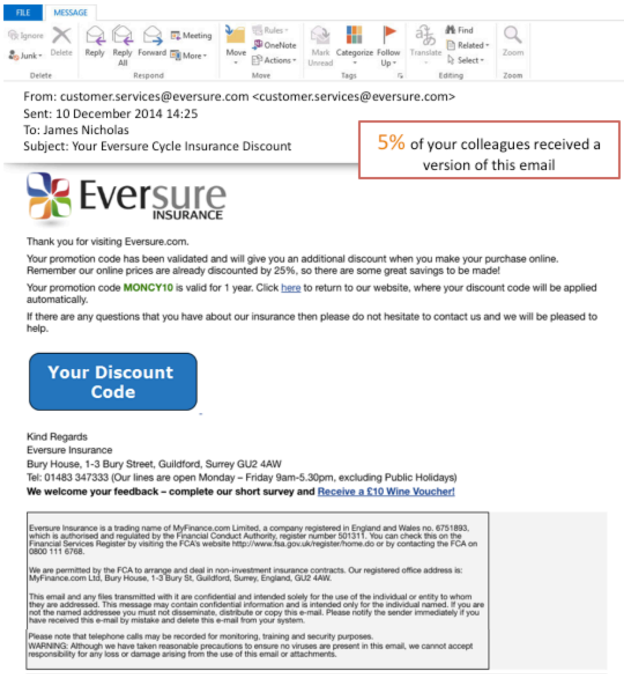
Additionally, we have updated the terms and conditions of our User Agreement that apply to your use of the products and services provided by Amazon Payments. Our updated User Agreement revises certain terms (including, among other things, the elimination of person-to-person payments). Our new User Agreement will become effective on October 13, 2014, which is more than 30 days from when we first posted our updated User Agreement. By continuing to use our services after October 13, 2014, you are agreeing to be bound by the terms and conditions of our new User Agreement. See User Agreement.

Please take a moment to review these changes which may also be found by clicking the User Agreement/Policies link on our web site at https://payments.amazon.com.

If you have questions or concerns about this information, please contact us by signing in to your Amazon Payments account and clicking on the Contact Us link here or by writing to us at Amazon Payments, Attn: Compliance, P.O. Box 81226 Seattle, Washington 98108-1226.

Thank you for using Amazon Payments.

Sincerely,
The Amazon Payments Team



From: customer.services@eversure.com <customer.services@eversure.com>
Sent: 10 December 2014 14:25
To: James Nicholas
Subject: Your Eversure Cycle Insurance Discount

5% of your colleagues received a version of this email

Eversure INSURANCE

Thank you for visiting Eversure.com.

Your promotion code has been validated and will give you an additional discount when you make your purchase online. Remember our online prices are already discounted by 25%, so there are some great savings to be made!

Your promotion code **MONCY10** is valid for 1 year. Click here to return to our website, where your discount code will be applied automatically.

If there are any questions that you have about our insurance then please do not hesitate to contact us and we will be pleased to help.

Your Discount Code

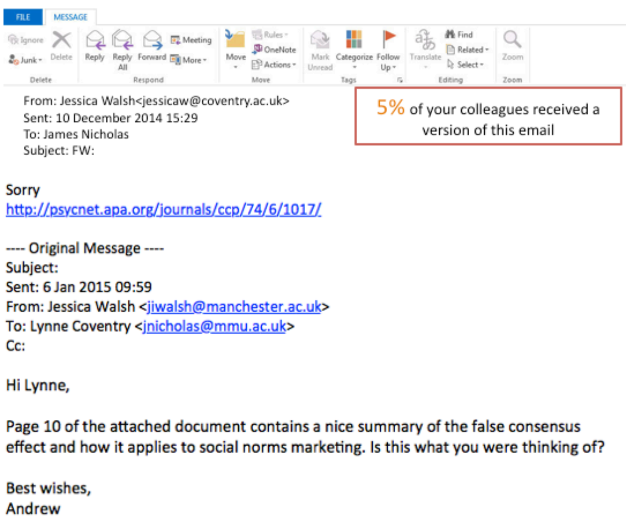
Kind Regards
Eversure Insurance
Bury House, 1-3 Bury Street, Guildford, Surrey GU2 4AW
Tel: 01483 347333 (Our lines are open Monday – Friday 9am-5.30pm, excluding Public Holidays)
We welcome your feedback – complete our short survey and Receive a £10 Wine Voucher!

Eversure Insurance is a trading name of MyFinance.com Limited, a company registered in England and Wales no. 6751893, which is authorised and regulated by the Financial Conduct Authority, register number 501311. You can check this on the Financial Services Register by visiting the FCA's website http://www.fsa.gov.uk/register/home.do or by contacting the FCA on 0800 111 6768.

We are permitted by the FCA to arrange and deal in non-investment insurance contracts. Our registered office address is: MyFinance.com Ltd, Bury House, 1-3 Bury St, Guildford, Surrey, England, GU2 4AW.

This email and any files transmitted with it are confidential and intended solely for the use of the individual or entity to whom they are addressed. This message may contain confidential information and is intended only for the individual named. If you are not the named addressee you must not disseminate, distribute or copy this e-mail. Please notify the sender immediately if you have received this e-mail by mistake and delete this e-mail from your system.

Please note that telephone calls may be recorded for monitoring, training and security purposes.
WARNING: Although we have taken reasonable precautions to ensure no viruses are present in this email, we cannot accept responsibility for any loss or damage arising from the use of this email or attachments.



From: Jessica Walsh<jessicaw@coventry.ac.uk>
Sent: 10 December 2014 15:29
To: James Nicholas
Subject: FW:

5% of your colleagues received a version of this email

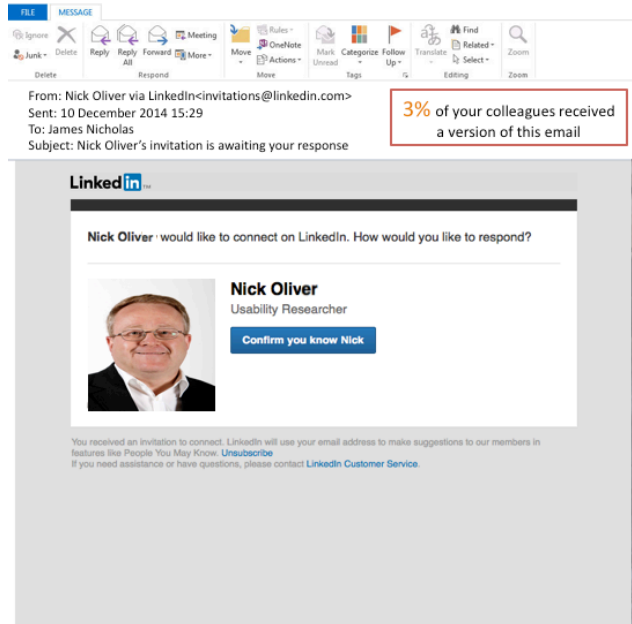
Sorry
http://psycnet.apa.org/journals/ccp/74/6/1017/

---- Original Message ----
Subject:
Sent: 6 Jan 2015 09:59
From: Jessica Walsh <jiwalsh@manchester.ac.uk>
To: Lynne Coventry <jnicholas@mmu.ac.uk>
Cc:

Hi Lynne,

Page 10 of the attached document contains a nice summary of the false consensus effect and how it applies to social norms marketing. Is this what you were thinking of?

Best wishes,
Andrew



From: Nick Oliver via LinkedIn<invitations@linkedin.com>
Sent: 10 December 2014 15:29
To: James Nicholas
Subject: Nick Oliver's invitation is awaiting your response

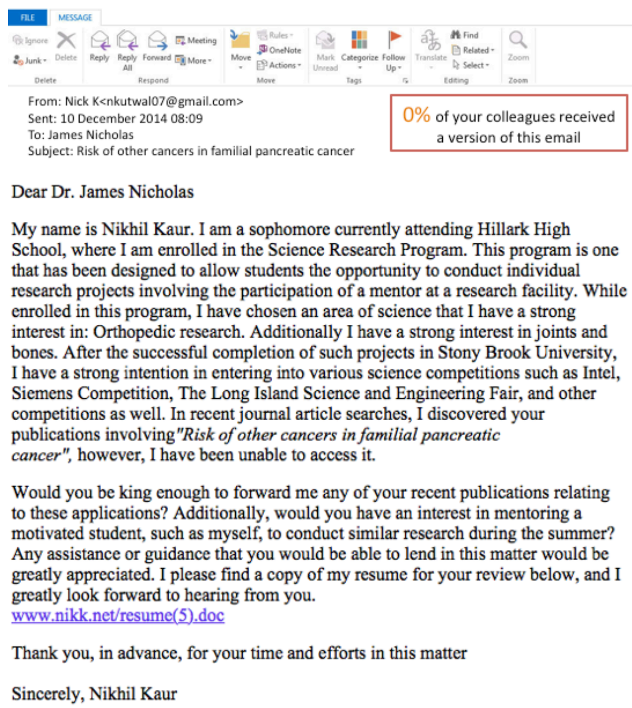
3% of your colleagues received a version of this email

Linked in

Nick Oliver would like to connect on LinkedIn. How would you like to respond?

Nick Oliver
Usability Researcher
Confirm you know Nick

You received an invitation to connect. LinkedIn will use your email address to make suggestions to our members in features like People You May Know. Unsubscribe
If you need assistance or have questions, please contact LinkedIn Customer Service.



From: Nick K<nkutwal07@gmail.com>
Sent: 10 December 2014 08:09
To: James Nicholas
Subject: Risk of other cancers in familial pancreatic cancer

0% of your colleagues received a version of this email

Dear Dr. James Nicholas

My name is Nikhil Kaur. I am a sophomore currently attending Hillark High School, where I am enrolled in the Science Research Program. This program is one that has been designed to allow students the opportunity to conduct individual research projects involving the participation of a mentor at a research facility. While enrolled in this program, I have chosen an area of science that I have a strong interest in: Orthopedic research. Additionally I have a strong interest in joints and bones. After the successful completion of such projects in Stony Brook University, I have a strong intention in entering into various science competitions such as Intel, Siemens Competition, The Long Island Science and Engineering Fair, and other competitions as well. In recent journal article searches, I discovered your publications involving *"Risk of other cancers in familial pancreatic cancer"*, however, I have been unable to access it.

Would you be king enough to forward me any of your recent publications relating to these applications? Additionally, would you have an interest in mentoring a motivated student, such as myself, to conduct similar research during the summer? Any assistance or guidance that you would be able to lend in this matter would be greatly appreciated. I please find a copy of my resume for your review below, and I greatly look forward to hearing from you.
www.nikk.net/resume(5).doc

Thank you, in advance, for your time and efforts in this matter

Sincerely, Nikhil Kaur

Raise the Curtains: The Effect of Awareness About Targeting on Consumer Attitudes and Purchase Intentions

Sonam Samat
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15232
sonamsam@andrew.cmu.edu

Alessandro Acquisti
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15232
acquisti@andrew.cmu.edu

Linda Babcock
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15232
lb2k@andrew.cmu.edu

ABSTRACT

We investigate the effect of awareness about targeting on users' attitudes towards a targeted ad and behavioral intentions towards the advertised product. Specifically, we study the effect of a notice that makes individuals aware that a particular advertisement has been targeted to them on their attitudes about the product and intentions to purchase the product. We find that, among individuals who have negative opinions about the practice of targeted advertising, awareness about targeting significantly worsens attitudes towards the targeted product and reduces the likelihood of purchasing the targeted product. Among individuals who have positive and neutral opinions about targeted advertising, awareness about targeting does not impact attitudes or purchase intentions towards the targeted product. We develop a scale to measure opinions about targeted ads and find that a substantial proportion (at least 33%) of our participants have negative opinions about targeted ads. This suggests that the self-regulated advertising industry is not incentivized to follow recommendations from the U.S. Federal Trade Commission to make consumers aware about their targeted advertising practices.

1. INTRODUCTION

Behavioral or targeted advertising is defined as "the practice of tracking an individual's online activities in order to deliver advertising tailored to the individual's interests" [16]. These online activities include webpages visited and content viewed, search queries, and online purchases. Many of the steps involved in performing targeted advertising (data collection and aggregation, data mining, bidding for ad spaces, etc.) are hidden from consumers. Consumers are typically only asked for overall consent for data collection when they register with an online service. After this initial consent, there are rarely any reminders about the continuous practice of online data collection. Consumers are also not made aware of the aggregation of their data from multiple sources, the mining of their data to select ads shown to them, or the real-time bidding process used to sell ad spaces. There is no easy way for consumers to distinguish between targeted ads and non-targeted ads, or to figure out what information was used in the targeting process. Therefore, as far as consumers are concerned, targeted advertising happens "behind the scenes." We investigate the effect of "raising the curtains," by making individuals aware that a particular ad is targeted to them, on their attitudes and purchase intentions towards the advertised product.

The industry favors the use of targeted advertising because it generates higher click-through rates [14] and higher sales [2] in comparison to non-targeted ads. But consumer surveys about perceptions of targeted advertising suggest that, by and large, people do not like being tracked and do not wish to receive behaviorally targeted advertisements [29, 34, 39]. These concerns are not unfounded given the growing privacy risks associated with large-scale data collection and use, such as the use of consumer data for price discrimination [28, 40] and in revealing embarrassing personal details [20]. In order to address these privacy risks, the U.S. Federal Trade Commission (FTC) has laid out a number of recommendations for best practices, one of which focuses on being transparent about how consumer information is collected and used, so that consumers are not kept in the dark and they can make informed decisions about their online activities [15, 16, 17].

The ad industry has made some efforts to achieve this goal. In 2011, the Digital Advertising Alliance (DAA), which is a coalition of advertising, media, marketing and technology companies, developed a set of icons (Figure 1) that may be displayed on targeted advertisements delivered by its members [9, 11]. The goal of these icons is to communicate how behavioral targeting works and provide consumers with avenues for opting out. However, Leon et al.'s 2012 work finds that people severely misunderstand these icons: 53% of their participants incorrectly believed that more ads would pop up if they clicked on the icon and 45% incorrectly believed that the accompanying 'AdChoices' tagline was intended to sell advertising space [22].

Figure 1. Icons used by ad industry on targeted ads.

We are interested in understanding how awareness about targeted advertising impacts consumers' attitudes and purchase intentions towards the advertised product. Previous research provides contradictory evidence about this effect. Research from the recommendations systems literature suggests that providing explanations for how recommendations are selected increases users' trust in the recommendation system and their likelihood to

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12 – 14, 2017, Santa Clara, California.

use the system in the future [8, 37]. Awareness about targeting could work in a similar way if it helps users understand how advertisements are selected for them and if it increases users' purchase intentions towards the advertised product. On the other hand, highly personalized advertisements can be perceived as intrusive or creepy, prompting individuals to avoid the advertisement [12, 26, 41]. Awareness about targeting may elicit feelings of intrusiveness because it reminds people about the continuous tracking of their data, and may therefore decrease purchase intentions towards the advertised product. By understanding how awareness about targeting impacts consumers' attitudes and purchase intentions, we aim to inform policy makers about the incentives of a self-regulated ad industry in implementing the FTC's recommendation about being transparent when targeting.

To this end, we conducted a series of studies to understand the impact of awareness about targeting on consumers' attitudes and purchase intentions towards the advertised product. We implemented awareness through the use of a text message shown along with the ad that indicated that the displayed ad had been selected for the user based on some information about the user. We hypothesize that the effect of awareness about targeting on attitudes and purchase intentions depends on individuals' overall opinions about the practice of targeted advertising. To test this hypothesis, we built and validated a scale that measures individuals' overall opinions about targeted advertising, as existing literature has not established a scale to measure this construct. We find that, for the participants who have negative opinions about targeted ads, awareness about targeting worsens attitudes and reduces purchase intentions for the advertised product. We also find that a substantial portion of our participants have negative opinions about targeted ads (53% in our first study and 33% in our third study). This suggests that a self-regulated ad industry is not incentivized to use notices that make consumers aware of their targeting practices, as it can significantly reduce their revenues from advertising.

2. RELATED WORK

2.1.1 Recommendation Systems

Recommendation systems (or recommendation agents) like Netflix or Amazon share several similarities with the practice of targeted advertising. Recommendation systems also collect data about consumers' preferences and behaviors, and use this data to recommend products to consumers. Research from this field shows that explaining how a recommendation was selected generally has a positive effect on users' attitudes. For instance, Chen and Pu's 2005 work shows that explanations can be an effective way to increase users' trust in the recommendation system [8]. In their 2002 work, Sinha and Swearingen asked participants to report how much they like the recommendations made by music recommender systems as well as how transparent they perceive these systems to be. They found that users like the recommendations made by systems they perceive to be transparent more than those they perceive to be non-transparent [37]. Bilgic and Mooney's 2005 work further shows that some explanations can even improve users' own accuracy in predicting how much they will like the recommended item [4]. In addition, 86% of the participants in a survey conducted by Herlocker et al. in 1999 said they would like to see explanations for the choices made by the recommendation system used in their experiment [19]. Johnson and Johnson's 1993 work attempted to explain this positive effect by highlighting that explanations provide an association between antecedent and consequent, that is, a link between cause and effect [21]. Notices that make consumers aware of the fact that an advertisement is targeted to them may also provide an explanation for why the particular advertisement was selected for them and, similar to recommendation systems, could improve consumers' attitudes about the advertised product.

There are some key differences between recommendation systems and behavioral targeting. Users can actively subscribe to recommendation systems, whereas they are typically subjected to targeted ads without their explicit consent. Therefore, the extent to which individuals expect (or even desire) personalized recommendations may be different for recommendations systems and behavioral targeting. While users expect recommendations for movies on Netflix and consumer goods on Amazon, they may not expect information from their Google search queries to be used in targeted advertisements on the New York Times website. As a result, the positive effect of explanations on users' intentions may be restricted to recommendations made within a specific platform and may not carry forward to behavioral targeting of advertisements that we study in this paper.

2.1.2 Targeted Advertising

Awareness about targeting might positively influence attitudes towards the targeted product due to the placebo effect, by which simply telling someone a product has been selected for them can positively influence their opinions about the product. Cosley et al.'s 2003 work shows that users' ratings of a recommendation can be manipulated simply by showing them what the recommendation system predicts their rating of the product will be, irrespective of whether the prediction is accurate or not [10]. In another example, a recent study conducted by OkCupid.com (an online dating platform) shows that the effect of simply telling people that a particular individual is a 90% match for them, when truly the individual is only a 30% match, is just as strong as when the individual is actually a 90% match [35]. In other words, the mere myth of compatibility works just as well as actual compatibility. This may happen if consumers trust that sophisticated algorithms with large amounts of data will make the best selections for them. In our context, a placebo effect may lead to improved attitudes and higher purchase intentions toward the advertised product if awareness notices convey the fact that an ad has been *selected* or *recommended* for the user.

Additionally, consumers may react to what a targeted ad signals about themselves. For instance, upon viewing a targeted ad for an environmentally friendly product, a consumer might believe that the advertiser thinks of her as someone who is environmentally friendly (i.e., the targeted ad can act as an implied social label). She may then adjust her own perceptions about being environmentally friendly and in turn also change her behavior towards environmentally friendly products. In their 2016 work, Summers et al. demonstrate that targeted advertisements can act as social labels causing consumers to adjust their own self-perceptions, and even behaviors, to match the implied labels. While they do not directly test if making individuals aware of targeting impacts purchase intentions, they explore whether awareness notices act as implied social labels and cause adjustments to self-perceptions, which in turn influence purchase intentions towards the advertised product. [38]. We borrow from the study design used by Summers et al. in their 2016 paper and extend their work by testing the direct impact of awareness notices on attitudes and purchase intentions towards the advertised product. More importantly, we test whether this impact varies with the *a priori* opinions that individuals may have about targeted advertising.

Consumer surveys on perceptions toward behavioral targeting help provide support for the opposite hypothesis that awareness about targeting can negatively influence attitudes and purchase intentions. In a nationally representative survey conducted by Turow et al. in 2009, 66% of the participants claimed that they did not want behaviorally targeted advertisements [39]. In 2010, McDonald and Cranor found that the majority of their participants (55%) also did not wish to receive advertisements tailored to them [29]. A 2012 survey conducted by the Pew Research Center found that 68% of adult Internet users say that they are not okay with targeted advertising because they "don't like having [their] online behavior tracked and analyzed" [34]. These data suggest that attitudes and purchase intentions towards the targeted ad may be negatively affected if awareness notices inherently remind consumers about the continuous tracking of their personal information.

In their 2012 work, Leon et al. show that users struggle to use the existing tools for opting out of targeted advertising, including tools that block access to advertising websites, tools that set cookies indicating a user's preference to opt out of targeted advertising, and privacy tools that are built directly into web browsers [23]. If users experience lack of control over their data then awareness notices indicating which ads are specifically targeted to them may cause users to retaliate by specifically avoiding the products shown in these targeted ads. This retaliation is called 'reactance' [6], and has been shown to occur when ads are perceived to be highly intrusive [12] and when messages are perceived to be highly personalized [41].

3. Hypotheses

As highlighted in the previous section, existing literature does not provide a clear indication of how awareness about targeting may impact individuals' purchase intentions. We believe that consumers' opinions about the practice of targeted advertising will play a moderating role in determining how awareness about targeting impacts attitudes and purchase intentions towards the targeted product. Specifically, for consumers who have a positive opinion about targeted advertising in general, awareness about targeting will increase purchase intentions towards the advertised product relative to no awareness about targeting. On the other hand, awareness about targeting will decrease purchase intentions relative to no awareness about targeting for consumers who have a negative opinion about targeted advertising. These hypotheses are developed based on three factors that determine attitudes towards the product: 1) how much the person likes that product in the absence of targeting, 2) how useful the person finds having a product recommended to her based on her personal information (in other words, how much the person likes targeted advertising), and 3) how invasive the person finds targeted advertising (in other words, how much the person dislikes targeted advertising). When the individual is not aware that an advertisement is targeted, the second and third factors are not activated and only the first factor determines the individual's attitudes towards the product. However, when the individual is made aware of targeting, the second and third factors are activated and a combination of all three factors determines her attitudes towards the product. Holding the first factor constant, the relative strength of the second and third factors determines the influence of awareness about targeting on purchase intentions.

In order to test this moderating role, we need to effectively measure opinions about targeting practices. Existing literature has not established a scale to measure opinions about targeting practices. Several researchers have used ad-hoc measures for this construct, but none have attempted to build and validate a scale to measure it. For instance, Leon et al.'s 2012 work included four questions towards the end of their survey to measure participants' opinions towards behavioral advertising. These questions were: "How comfortable are you with behavioral advertising?", "In general, I find behavioral advertising useful", "In general, I like behavioral advertising" and "Behavioral advertising is privacy invasive" [22]. In 2016, Melicher et al. conducted in-person interviews with 35 participants asking them questions to capture their opinions about online tracking. They identified four groups of opinions (generally negative, generally neutral, mixed, and conditionally positive) but they did not attempt to build a scale to measure opinions in a closed-ended format [30]. In 2010, McDonald and Cranor interviewed 14 participants, and used their responses to create two closed-ended measures: one for feelings towards current targeting practices and another for reasons to accept or reject targeted advertising. They later used these measures in an online survey but they did not perform any formal validation analysis for their measures [29].

4. OUR CONTRIBUTION

To test our hypotheses, we conducted three studies. The first was an exploratory study to measure participants' opinions about targeted advertising using open-ended questions. The goal of this study was to capture the different kinds of opinions that participants express about targeting practices in their own words, and use them to build a scale that can measure opinions about targeting in future studies. The second study was a validation study in which we validated the closed-ended scale created from the responses received in the exploratory study. We also shortened the scale so participants can complete it quickly. We then tested convergent and divergent validity, as well as the test-retest validity of our scale. Finally, in the third study, we evaluated the effect of awareness about targeting on attitudes and purchase intentions towards the advertised product. In this study, we tested the hypothesis that opinions about targeting moderate the effect of awareness of targeting on attitudes and purchase intentions.

We recruited participants for all studies from Amazon Mechanical Turk. The platform allows researchers to approve or reject participants' payment based on their performance. Therefore, each participant has an approval rating (the percentage of his or her previously completed surveys or tasks that have been approved). Following the recommendations established by previous researchers, we implemented a minimum requirement of a 95% approval rating during our recruitment process [33]. All our studies were approved by the Institutional Review Board at Carnegie Mellon University and participants in all studies consented to participate in this research.

4.1 Exploratory Study

4.1.1 Methods

We believe consumers' opinions about targeting practices will play a moderating role in the effect of awareness on purchase intentions. The primary goal of this study was to identify the different types of opinions (for example: positive, negative, conditionally positive, etc.) that individuals have towards targeted advertising practices, and build a scale to measure these opinions reliably and quickly. From this study, we also gained insights into the prevalence of different opinions, the reasons behind these opinions, and the factors that correlate with these opinions.

We conducted an online survey on Amazon's Mechanical Turk platform. The survey lasted 15–20 minutes and participants were paid $1 as compensation. Participants were shown a hypothetical scenario about a targeted advertising experience. Targeting can be done in several different ways and individuals' opinions may vary across different types of targeting. In this study, we measured opinions about three types of targeting: 1) Remarketing: where the targeted ad is for a specific product that the individual has looked at before, 2) Interest-based targeting: where the targeted ad is for a product that is similar to other products that the individual has looked at before, and 3) Contextual targeting: where the targeted ad is related to the content of the website where the ad is shown. By random assignment, each participant in our study was shown one of three hypothetical scenarios involving one of the three types of ads (see Appendix A.1). Once participants read the scenario, they were asked open-ended questions to capture their opinions, and the reasons behind those opinions, about the type of targeting practice in their scenario. Specifically, we asked participants the following questions: Q1: "How would you feel if you experienced this scenario?" and Q2: "Please tell us why you would feel this way if you experienced this scenario." Finally, we measured individual-level factors such as the IUIPC scale for privacy concerns [27], previous experiences with online privacy invasions involving targeted ads and in general (self-developed), previous actions taken to avoid targeted ads [22], perceived control over this type of targeted advertising (self-developed), level of Internet usage [24], current interest in buying the advertised product (self-developed), the Domain-Specific Risk-Taking (DOSPERT) scale [5], Internet usage levels, and demographics. The self-developed measures are reported in Appendix A.2.

4.1.2 Results

One hundred eighteen participants (Mean Age = 31.4; 58% Male) completed this study. Two independent coders read and coded participants' open-ended responses about their opinions. After they completed the first round of coding independently, they met and discussed their codes to consolidate them and form the final set of codes. Then, they coded the open-ended responses once again using this final set of codes. The inter-rater reliability for the final codes was satisfactory for both questions (Cohen's kappa=0.71 for Q1 and Cohen's kappa=0.77 for Q2). The cases where the coders differed in their final codes (25% in Q1 and 19% in Q2) were resolved by the first author, by reading the participants' responses and selecting the code that seemed a better fit from among the two codes assigned by the coders.

4.1.2.1 Categories of Opinions

Eight different categories of opinions about targeted advertising emerged from participants' open-ended responses to Q1. Figure 2 shows the distribution of these opinions across all three conditions. As can be seen from this figure, a large proportion of our participants (37%) reported that they would feel neutral if they experienced the targeted advertising scenario shown to them. Some responses in this category were "I'd feel fine, that's pretty normal," "I would have neutral feelings," and "I would not feel any special way." A small proportion of our participants (6%) reported feeling positively about the targeted advertising scenario shown to them, and an even smaller proportion (4%) reported feeling mixed emotions. Some responses from the positive category were "[I] would feel interested" and "I would feel excited to have this new [recommendation] and I wouldn't really care that my browsing history was tracked," and a response we observed in the mixed category was "I would have mixed emotions. I'd feel a bit weird about the tracking thing, but I'd also be interested in checking out the shoes, probably."

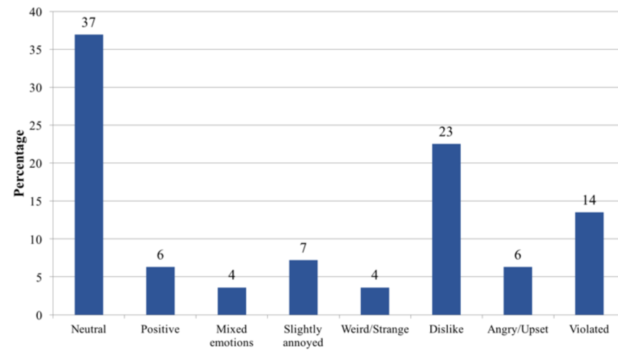


Figure 2. Distribution of participants' open-ended responses to Question 1.[1]

We observed a wide range of negative opinions towards these types of targeted ads. The spectrum of responses ranged from "slightly annoyed" to "violated," with three categories in between these two extremes. The total proportion of participants that reported negative feelings is 53%. Some examples of responses from the modal 'Dislike' category were "I would feel like my personal information is not safe. It would make me feel uncomfortable," "I would feel watched and unhappy," and "I would feel like my privacy has been invaded."

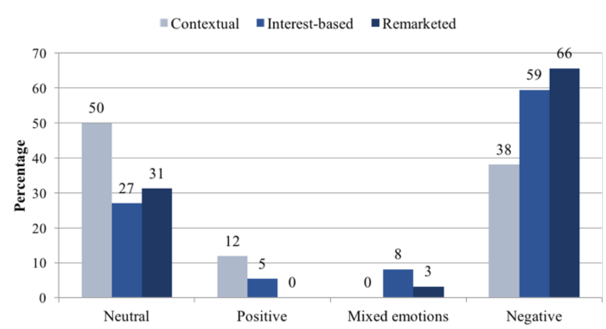


Figure 3. Distribution of participants' open-ended responses to Question 1, split by targeting type condition.

Next, we looked at the differences between the three conditions (contextual, interest-based, and remarketed). For simplicity, we combined the five categories of negative responses. As Figure 3 shows, the contextual ad condition had more neutral responses and

[1] Seven participants' responses to Q1 were coded as unusable either because the response did not make any sense or because the response did not answer the question in anyway. The percentages and graphs shown here are for the remaining 111 participants.

fewer negative responses than the interest-based and remarketed conditions. Further statistical tests confirmed that the distribution of responses (using the consolidated 4 categories) did not significantly differ from each other in the interest-based and remarketed conditions (χ^2 (3) = 2.66, p=0.44). However, the distribution of responses in the contextual condition was significantly different from the interest-based (χ^2 (3) = 8.86, p=0.031) and remarketing conditions (χ^2 (3) = 9.40, p=0.024).

4.1.2.2 Reasons Behind Opinions

Nine different categories emerged from participants' open-ended responses to Q2 about why they would feel the way they reported feeling in Q1. As shown in Figure 4, the most common reason we observed (provided by 29% of our participants) was "because it happens all the time," suggesting that individuals have become used to seeing targeted ads. Some responses we observed in this category were "I have had it happen before many times, so it has become normal," "I would feel this way because it is a scenario that I have experienced in the past, and currently experience," and "I would feel this way because it is not the first time this has happened. It's common practice in my opinion." Not surprisingly, most of these participants (87%) reported feeling neutral in Q1.

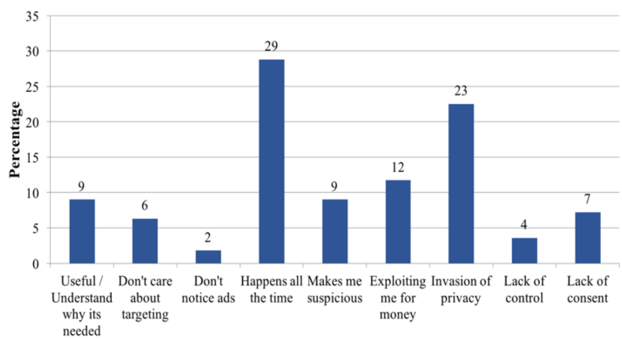


Figure 4. Distribution of participants' open-ended responses to Question 2.[2]

The second most common reason we observed (provided by 23% of our participants) was "because it feels like an invasion of my privacy." Many participants in this category used the exact words "because it feels like an invasion of my privacy" as part of their response, while some others reported "This would make me [feel] that [it's] not safe for me to look at anything online because it [feels] as if someone is watching me" and "they made me feel as if [I] was being watched." Almost all participants in this category (96%) reported feeling one of the negative opinions in Q1.

The third most common category we observed (provided by 12% of our participants) was "because I feel they are exploiting me for money" which includes responses where participants specifically said they feel they are being taken advantage of, or being manipulated for money, in such scenarios. Interestingly, 62% of participants who reported this reason were in the contextual targeting condition. One participant in the contextual targeting condition reported, "The advertiser pays them, so the newspaper feels obligated to place the ad somewhere that readers can feel [encouraged] to buy this particular shoe. They are trying to lead unsuspecting readers to this particular store. It's consumer manipulation." Other categories we observed were "because I find such ads useful" or "because I understand why they are needed" (9%), "because it makes me suspicious" (9%), "because I did not provide permission" (7%, labeled 'lack of consent' in Figure 16), "because I don't care about targeting" (6%), "because I do not know how to turn it off" (4%, labeled 'lack of control' in Figure 17), and "because I don't notice ads" (2%).

4.1.2.3 Correlations with Opinions

Next, we explored how perceived control over targeted advertising and previous experiences with online privacy invasions correlate with opinions. Before doing this analysis, we noted that the biggest category of responses is negative and that the other categories are relatively small, suggesting that consolidating categories might be necessary. Accordingly, we created two consolidated categories for 'negative' vs. 'non-negative' opinions, combining the neutral, mixed, and positive categories into one. We ran logistic regressions with the consolidated two-category dummy variable for negative opinions as our dependent variable (coded as '0' for non-negative opinions and '1' for negative opinions) and the individual-level factors measured in our survey as the independent variables. This analysis suggests that, for the interest-based and remarketing conditions, the level of perceived control over targeting significantly predicts participants' opinions about targeting (odds ratio=0.30, p=0.002). Participants in the interest-based and remarketing conditions who feel more control over this type of targeting are less likely to have negative opinions about it. In the contextual targeting condition, a previous unpleasant experience with targeted ads marginally predicts opinions, such that participants who report having a previous unpleasant experience with targeted ads are more likely to have negative opinions about targeting (odds ratio = 7.7, marginally significant p = 0.054). The full results of our logistics regressions are shown in Appendix B.1. We validate these relationships with a larger sample size in Validation Study A.

4.1.3 Discussion

Our results suggest that many participants (53%) feel negatively about targeted advertising, especially interest-based and re-marketed advertising. It may be that the collection and use of browsing history (which happens in the interest-based and remarketing scenarios but not in the contextual targeting condition) plays a key role in explaining how participants feel about targeting practices.

In this study, we also learn that a substantial number of participants (37%) have neutral opinions about targeting practices. Given this result, we re-evaluate our original hypothesis about the moderating role of opinions about targeting in the effect of awareness on purchase intentions. What would be the effect of awareness about targeting on purchase intentions for individuals who feel neutral about targeting practices? As 68% of participants who reported having neutral opinions also reported feeling this way because targeting "happens all the time," we believe there will be no effect of awareness notices on purchase intentions for these individuals.

[2] Seven participants' responses to Q2 were coded as unusable either because the response did not make any sense or because the response did not answer the question in any way. The percentages and graphs shown here are for the remaining 111 participants.

These individuals expect such targeting to occur anyway, so they will obtain no new information from awareness messages. We find that a small proportion of individuals reported positive or mixed opinions about targeted ads. While the effect of awareness about targeting may be different for these individuals, it is difficult to test such an effect if very few people report having such opinions.

In the next studies, we narrow our focus to develop a scale to measure opinions about one type of targeted advertising because different types of targeting may require different measurement models. We choose to focus on interest-based targeting because we believe it is more difficult for individuals to identify interest-based ads in comparison to remarketed and contextually targeted ads, since remarketed ads advertise the same product that an individual has looked at before and contextual ads advertise a product related to the content of the page where the ad is shown. In comparison, interest-based ads display products similar, though not identical, to ones that an individual has looked at before and as such the consumer may not be aware of the targeting.

4.2 Validation Study A

4.2.1 Methods

In order to measure opinions about the practice of targeted advertising without having to collect and code open-ended responses, we built a scale starting with the 14 most common responses provided by our participants in the first question in the Exploratory Study (Table 1). In this study, we recruited participants from Amazon's Mechanical Turk platform for a study advertised to collect individuals' opinions about a hypothetical online scenario. In order to avoid selection bias, our Mechanical Turk post did not mention anything about targeted ads (or ads in general). The survey lasted 5 minutes and participants were paid $0.30 as compensation. Participants were shown the interest-based targeting scenario used in the Exploratory Study. After participants read the scenario, they were told, "We are interested in understanding how you would feel about the type of targeted advertising described in the scenario on the previous page. Please indicate how strongly you agree or disagree with the following statements." For each statement, participants were asked to indicate their level of agreement on a 1–7 scale from 'Strongly Agree' to 'Strongly Disagree'. The order of the 14 statements was randomized.

We asked participants their perceived level of control over this type of targeted advertising (self-developed) and their previous experiences with online privacy invasions either involving targeted ads or in general (self-developed). Then, we asked questions about their current interest in purchasing shoes (the product used in our ad), how often they purchase shoes online and offline, their general Internet usage [25], actions they have taken to avoid being tracked online [22], their Internet privacy concerns [27], and demographics.

4.2.2 Results

Two hundred ninety-six participants (Mean Age = 34.3; 53% Male) completed this study. Our 14-item scale shows high reliability (Cronbach's alpha = 0.9265, well above the acceptable threshold of 0.70 [32]). As can be seen in Table 1, eliminating the two mixed category items results in a higher Cronbach's alpha value of 0.9554. Therefore, we decided to drop the two mixed items from our scale.

Table 1. Correlations, covariance, and Cronbach's alpha values for the 14-item scale. *Cronbach's alpha value only increases when the two mixed category items are deleted.

Item	Category	Item-test correlation	Item-rest correlation	Average inter-item covariance	Alpha if item is deleted
I would feel ok about this type of advertising	Neutral	0.90	0.88	1.49	0.9141
I would feel indifferent about this type of advertising	Neutral	0.63	0.56	1.60	0.9246
I would not be bothered by this type of advertising	Neutral	0.83	0.80	1.51	0.9168
I would think this is clever advertising	Positive	0.68	0.62	1.58	0.9228
I would be interested in products shown in this type of advertising	Positive	0.76	0.72	1.55	0.9195
I would feel this type of advertising can be helpful to me	Positive	0.81	0.78	1.54	0.9178
I would have mixed feelings (good and bad) about this type of advertising	Mixed	0.08	-0.03	1.83	0.9428*
This type of advertising would make me feel worried but I can also see the benefit to me	Mixed	0.14	0.05	1.80	0.9392*
I would be annoyed by this type of advertising	Negative	0.87	0.84	1.50	0.9154
I would feel creeped out by this type of advertising	Negative	0.86	0.83	1.48	0.9155
I would not like this type of advertising	Negative	0.88	0.86	1.49	0.9148
This type of advertising would make me feel like my privacy has been invaded	Negative	0.86	0.83	1.50	0.9157
I would feel upset about this type of advertising	Negative	0.83	0.79	1.51	0.9170
This type of advertising would make me feel violated and manipulated	Negative	0.87	0.85	1.49	0.9152
Overall				1.56	0.9265

Table 2. The 6 items in our final scale, along with their inter-item correlations. *Significant at the 0.001 alpha level.

	Item	Category	1	2	3	4	5	6
1	I would feel ok about this type of advertising	Neutral	1					
2	I would not be bothered by this type of advertising	Neutral	0.76*	1				
3	I would be interested in products shown in this type of advertising	Positive	0.70*	0.62*	1			
4	I would feel this type of advertising can be helpful to me	Positive	0.74*	0.65*	0.76*	1		
5	I would be annoyed by this type of advertising	Negative	-0.79*	-0.71*	-0.63*	-0.65*	1	
6	I would not like this type of advertising	Negative	-0.79*	-0.71*	-0.64*	-0.71*	0.78*	1

4.2.2.1 Scale Consolidation

Our next goal was to reduce the size of the scale so participants can complete it in a shorter period of time, while still maintaining the high reliability of the scale. We decided to reduce the scale to 6 items, and consequently computed the Cronbach's alpha value for all combinations of 6 items from the remaining 12 items. We identified the best combination as the one that equally represents the remaining three categories of opinions (neutral, positive, and negative) and has a high Cronbach's alpha value of 0.9362. Table 2 shows the 6 items that make our final scale. All pairwise correlations between the 6 items are statistically significant at the 0.001 alpha level. In addition to reducing the time burden on participants, this shorter scale also ensures that it is not biased towards any single category as it includes an equal number of items (two) from each category.

Table 3. Factor loadings from the exploratory factor analysis (Validation Study A) and confirmatory factor analysis.

Item	Factor Loading (EFA)	Completely Standardized Parameter (CFA)
I would feel ok about this type of advertising	0.9021	0.9378
I would not be bothered by this type of advertising	0.8115	0.8183
I would be interested in products shown in this type of advertising	0.7890	0.8513
I would feel this type of advertising can be helpful to me	0.8311	0.8957
I would be annoyed by this type of advertising	-0.8456	-0.8528
I would not like this type of advertising	-0.8640	-0.8406

4.2.2.2 Exploratory Factor Analysis

Exploratory factor analysis with our 6-item scale yielded a single factor with eigenvalue greater than one. As shown in Table 3, each of the 6 items is highly correlated with the single underlying factor. This suggests that our scale measures a single underlying construct: opinions towards targeted advertising. Positive and neutral items have positive factor loadings while the negative items have negative factor loadings. Therefore, our scale can be consolidated into a single value by reverse coding the negative items (8 – value) and then averaging the 6 items. Lower numbers of the consolidated value indicate more negative opinions towards targeted advertising. We use this consolidated value to validate the preliminary results obtained in the Exploratory Study. Just as in the Exploratory Study, we find that the level of perceived control over targeting significantly predicts participants' opinions about targeting ($\beta_{control} = 0.140$, p=0.011). We also find that having a previous unpleasant or uncomfortable experience with targeted ads significantly predicts participants' opinions about targeted ads ($\beta_{prevtar} = -1.16$, p<0.001). The regression coefficients are reported in Appendix B.2.

4.3 Validation Study B

4.3.1 Methods

The goal of this study was three-fold: 1) to confirm the single-factor model for our 6-item scale that emerged in Validation Study A, 2) to test the divergent validity of our 6-item scale, and 3) to evaluate the test-retest reliability of our 6-item scale. Towards the first goal, we recruited participants from Amazon Mechanical Turk in the same way as in Validation Study A (participants who completed Validation Study A were not allowed to take this survey). The survey lasted 10–15 minutes and participants were paid $0.75 as compensation. Participants were shown the hypothetical scenario about interest-based targeting and asked to indicate their level of agreement with the 6 items (presented in random order) on a 1–7 scale from 'Strongly Agree' to 'Strongly Disagree'.

Towards the second goal for this study, we included three existing scales in our survey: 1) the IUIPC privacy concerns scale [27], which measures Internet users' information privacy concerns, 2) the Domain-Specific Risk-Taking (DOSPERT) scale [5], which measures individuals' attitudes towards engaging in risky behaviors, and 3) the General Decision Making Style (GDMS) scale [36], which measures decision-making styles. Recent work

Table 4. Scale descriptive statistics, reliabilities, and correlations with our 6-item scale. The three sub-scales of the IUIPC scale show significant (at the 0.001 alpha level) and moderate correlations with our 6-item scale.

Scale	Sub-scale	Mean	Standard deviation	Cronbach's alpha	Correlation with our 6-item scale
IUIPC	Control	5.68	1.08	0.7903	-0.36*
	Awareness	6.03	1.06	0.8654	-0.25*
	Collection	5.44	1.34	0.9095	-0.47*
DOSPERT	Financial	2.72	1.36	0.8577	0.07
	Health/Safety	2.83	1.22	0.7261	0.08
	Recreational	2.88	1.45	0.8392	0.01
	Ethical	2.24	1.10	0.8097	0.05
	Social	4.81	1.18	0.7779	-0.09
GDMS	Rational	3.97	0.66	0.8330	-0.07
	Avoidant	2.43	1.06	0.9227	0.01
	Dependent	3.07	0.84	0.8329	0.06
	Intuitive	3.09	0.87	0.8550	0.09
	Spontaneous	2.44	0.90	0.8620	0.11

has shown that these three established scales are good predictors of privacy preferences and behaviors [13]. Towards the third goal, all participants were invited to take a follow-up survey after a two-week gap, in which they were shown the same scenario and asked to indicate the extent to which they agreed with the same 6 items.

4.3.2 Results

Two hundred ninety-four participants (Mean Age = 33.8; 53% Male) completed this study. Again, our 6-item scale shows high reliability (Cronbach's alpha = 0.9385) and the elimination of no item yields a higher Cronbach's alpha (α-if-item-deleted$_i$ < 0.9385, for all items).

4.3.2.1 Confirmatory Factor Analysis

Confirmatory factor analysis suggests that the data fit a single-factor model well, with a Bentler's comparative fit index (CFI) of 0.955, a Tucker-Lewis index (TFI) of 0.925, and a standardized root mean squared residual (SRMR) of 0.027. The confirmatory factor loadings are shown in Table 3 and are all significant at the 0.001 alpha level with t-statistics greater than 29.5.

4.3.2.2 Divergent Validity

Next, we tested the divergent validity of our 6-item scale. We compared our scale to three established scales: IUIPC, DOSPERT, and GDMS. Table 4 shows the mean, standard deviations, Cronbach's alpha coefficients, and correlations with our 6-item scale for each subscale of the three established scales. The IUIPC and DOSPERT scales use a 1–7 response scale while the GDMS scale uses a 1–5 response scale. As shown in Table 5, only the sub-scales of the IUIPC scale are significantly correlated with our 6-item scale (at the 0.001 alpha level). Correlations of our 6-item scale with the sub-scales of the DOSPERT and GDMS scales are all not significant. The significant correlations between our 6-item scale and the three sub-scales of the IUIPC scale are negative and moderate in nature. It makes intuitive sense that peoples' opinions about targeted ads are correlated with their informational privacy concerns, but given the moderate size of these correlations, we can conclude that our measure is distinct from the construct measured by the IUIPC scale.

We conducted a second test of divergent validity by comparing the square root of the Average Variance Extracted (AVE) by our single factor with the correlations between our 6-item scale and the three established scales. The square root of the AVE by our single factor is 0.849, which is higher than all the correlations between our 6-item scale and the three established scales, thus establishing discriminant validity.

4.3.2.3 Test-Retest Reliability

Next, we evaluated the test-retest reliability of our 6-item scale. One hundred sixty-six of the original two hundred ninety-four participants completed our follow-up survey (response rate = 57%). The survey lasted 5 minutes and participants were paid \$0.30 as compensation. Our 6-item scale continues to show high internal reliability (Cronbach's alpha = 0.9383) and also shows high test-retest reliability ($r(164) = 0.75$, $p<0.001$). Therefore, our 6-item scale is validated to reliably measure opinions about the practice of targeting advertising.

4.3.3 Discussion

In Validation Studies A and B, we validated our 6-item scale to measure opinions about targeted ads. Our scale shows high reliability, test-retest validity, and discriminant validity. Therefore, our 6-item scale can be used to reliably measure individuals' opinions about targeted ads. In order to use our scale, researchers should present participants with the hypothetical scenario about interest-based targeting shown in Appendix A.1. Then, participants should be told, "We are interested in understanding how you would feel about the type of targeted advertising described in the scenario on the previous page. Please indicate how strongly you agree or disagree with the following statements." The 6 items should be presented in random order and, for each item, participants should be asked to indicate their level of agreement on a 1–7 scale from 'Strongly Agree' to 'Strongly Disagree'.

4.4 Evaluation Study

4.4.1 Methods

The goal of this study was to evaluate how awareness about targeting impacts participants' purchase intentions. We borrowed the methodology used by Summers et al. in their 2016 work but adapted their study design to meet the goals of our study [38]. Specifically, we first ran a pilot study in which participants evaluated the perceived environmental friendliness of 32 different products belonging to 8 different product categories (such as light bulbs, laundry detergents, notebooks, etc.). These 32 products were the same ones used by Summers et al. in their 2016 work and are listed in Appendix A.3. The goal of this pilot study was to confirm that the Mechanical Turk population perceives at least some of these 32 products to be environmentally friendly. Participants were shown the 32 products in random order (one product at a time) and asked to indicate on a 1–7 scale (1 – Not at all, 7 – A lot) how environmentally friendly each product seems to them relative to other products in the same product category. Next, we recruited new participants from the same population for our main study. This study was conducted in two phases. In the first phase, we captured participants' opinions towards targeting practices using our validated 6-item scale from the Validation Study. By measuring these opinions *before* we showed participants our ad stimulus, we were able to collect an unbiased measure of our hypothesized moderator variable. In this phase, we also measured participants' tendency to express their value of environmental protection through the purchase of goods and services by asking participants to complete the Green Consumption Values scale [18]. We included a third unrelated scale in phase 1, the 'Dysfunctional Beliefs and Attitudes about Sleep' scale [31], as a decoy in order to ensure that participants are not able to make an obvious guess about the goal of our study or draw a direct connection between the two phases of our study. Finally, participants were asked to answer some demographic questions.

After a gap of about two weeks, we invited the participants who completed phase 1 to an ostensibly new study, which constituted the second phase of our main study. This second phase was almost identical to the study conducted by Summers et al. in their 2016 work [38]. Participants were told they would need to complete three different tasks in this study. The first task was a shopping task in which participants were asked to select one product from a set of four product options in several different product categories (the list of products is provided in Appendix A.3). We used this task to make the story of targeted advertising plausible. The second task was an advertisement evaluation task in which participants were shown an advertisement for an acoustic speaker. Between conditions, we manipulated whether participants were made aware that the ad was targeted to them. We did this by telling participants in the 'Awareness' condition, "Our software will customize an advertisement for you based on your responses from the shopping task you completed earlier" before they were shown the ad. Participants in the 'No Awareness' condition were shown the same ad and were not given any information about targeting. In addition to the message about software matching, participants in the 'Awareness' condition were also shown the AdChoices icon on the acoustic speaker advertisement, the title of the page where the ad was shown was 'Targeted Advertisement', and the following text was shown above the ad: "The ad displayed below is customized for you based on your shopping choices earlier in today's session. Please take a moment to consider the advertisement below." In the 'No Awareness' condition the advertisement did not include the AdChoices icon, the title of the page was 'Advertisement', and the following text was shown above the ad: "Please take a moment to consider the advertisement below." All participants were shown the same advertisement, so it was not truly matched to any shopping choices. The advertisement shown to participants is provided in Appendix A.4.

As in Summers et al.'s 2016 work, we first asked participants how much they liked the advertisement on a 7-point scale (1 – Not at all, 7 – A lot) [38]. This question was included to confirm that participants' perceptions about the *advertisement* did not differ between conditions. Next, we asked participants how much they liked the product on a 7-point scale (1 – Not at all, 7 – A lot). This question captures attitudes towards the product. Finally, we measured purchase intentions towards the advertised product by using the same question as Summers et al. (2016). Specifically, we asked participants how likely they were to buy the acoustic speaker on a 7-point scale (1 – Very Unlikely, 4 – Undecided, 7 – Very Likely). On the next page, as a manipulation check, we asked participants the extent to which they agreed with the statement, "The advertisement shown to me was matched to my purchase choices from the earlier task in this study" on a 7-point scale (1 – Strongly Disagree, 7 – Strongly Agree).

Then participants continued to the final task in this study, which was almost identical to the final task used by Summers et al. in their 2016 work [38]. In this task, participants were informed that the researchers conducting this study have decided to partner with a different charity each month, to help our participants support these charities. They were informed that the selected charity this month is 'Rainforest Alliance' which "is a non-governmental organization (NGO) working to conserve biodiversity and ensure sustainable livelihoods by transforming land-use practices, business practices and consumer behavior." Participants were then informed that they would be entered in a lottery in which five participants would be randomly selected to win $10, and they can choose to donate some or all of their winnings to the Rainforest Alliance if they are selected as one of the winners. They were also informed that the researchers would match any donation they chose to make towards the Rainforest Alliance. Then, we asked participants if they would like to donate to the Rainforest Alliance should they be chosen as a winner in the lottery. Those who responded affirmatively were asked for the exact amount that they wished to donate. This measure of donation behavior was used as an additional dependent variable to test whether the impact of awareness about targeting carries forward to subsequent decisions related to the factor purportedly used in the targeting process (environmental friendliness). In other words, if participants feel that they were shown a targeted ad for an environmentally friendly product because the advertiser believes they are environmentally friendly, then they might change their behavior when deciding to subsequently donate to an environmentally friendly charity. Finally, participants were debriefed about the fact that the advertisement they saw was not truly targeted to them.

4.4.2 Results

4.4.2.1 Pilot Study

One hundred forty-four participants (Mean Age = 34.5; 60% Male) completed this study. The survey lasted 10 minutes and participants were paid $0.50 as compensation. In order to determine whether a product is perceived to be environmentally friendly, we tested whether the mean value of reported environmental friendliness was significantly different from the midpoint of the scale, 4. In six of the eight product categories, at least one of the four products was perceived to be environmentally friendly. In the remaining two product categories (mouthwashes and digital cameras), none of the four products was perceived to be environmentally friendly. In order to improve the plausibility of our targeting scenario, we dropped these two product categories from the shopping task in the second phase. Therefore, the shopping task in the second phase included four product options in each of the following six product categories: light bulbs, laundry detergents, notebooks, air purifiers, dish scrubbers, and water bottles. The full results of this pilot, along with the means and standard deviations of the similar pilot study conducted by Summers et al. in their 2016 work, are presented in Appendix B.3.

4.4.2.2 Phase 1

Nine hundred ninety-two participants (Mean Age = 35.7; 53% Male) completed this study. The survey lasted 10 minutes and participants were paid $0.50 as compensation. We created a consolidated value for opinions towards targeted ads by reverse coding the two negative items (8 – value) and then averaging the 6 scale items. Figure 5 shows the distribution of this consolidated value. Lower numbers indicate negative opinions towards targeted ads and higher numbers indicate neutral and positive opinions towards targeted ads. As is evident from Figure 5, many participants fell in the middle of the scale for our consolidated value of opinions about targeted ads. This may be because of participants' tendency to "pile on the midpoint" of response scales [1].

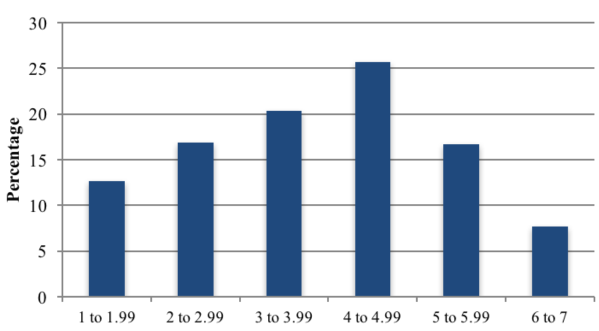


Figure 5. Distribution of opinions towards targeted ads using our 6-item scale.

4.4.2.3 Phase 2

Of the nine hundred ninety-two participants invited to participate in this study, six hundred ninety-seven (Mean Age = 36.7; 55% Male) completed Phase 2 (Response Rate = 70.3%). The survey lasted 10 minutes and participants were paid $2 as compensation. The full results of all the models reported below in this section are presented in Appendix B.4. First, we analyze the responses to our manipulation check questions. Participants in the 'Awareness' condition believed that the advertisement was matched to their purchase choices more than those in the 'No Awareness' condition ($Mean_{Awareness}$ = 4.68, $Mean_{NoAwareness}$ = 3.54; t = 8.78; p<0.001). This suggests that participants in the 'Awareness' condition believed that the advertisement was targeted. Participants' attitudes towards the advertisement, measured by how much they like the ad, were not significantly different between the two conditions ($Mean_{Awareness}$ = 4.46, $Mean_{NoAwareness}$ = 4.55; t = 0.84; p=0.40). Next, we analyze the responses for our dependent variables, attitudes and purchase intentions. Overall, we find no significant effect of awareness about targeting on attitudes towards the product ($Mean_{Awareness}$ = 4.56, $Mean_{NoAwareness}$ = 4.65; t = 0.63; p=0.53) or purchase intentions towards the product ($Mean_{Awareness}$ = 3.52, $Mean_{NoAwareness}$ = 3.47; t = 0.37; p=0.71).

Next, we separately analyze how awareness about targeting impacts attitudes and purchase intentions towards the product among 1) participants with negative opinions about targeted advertising and 2) participants with neutral and positive opinions about targeted advertising. The former category comprises participants whose consolidated value on our measure for opinions about targeted advertising is 3 or lower (N=228), and the latter category comprises participants whose consolidated value on our measure for opinions about targeted advertising is 5 or greater (N=176). We estimate the following econometric models separately for these two categories of participants:

$$LikeProduct_i = \beta_0 + \beta_{Awareness}\,Awareness + \beta_{Male}\,Male + \beta_{Age}\,Age + \beta_{Caucasian}\,Caucasian + \beta_{Education}\,Education + \varepsilon_i$$

$$BuyProduct_i = \beta_0 + \beta_{Awareness}\,Awareness + \beta_{Male}\,Male + \beta_{Age}\,Age + \beta_{Caucasian}\,Caucasian + \beta_{Education}\,Education + \varepsilon_i$$

where 'LikeProduct' and 'BuyProduct' represent participants' attitudes towards the product and purchase intentions towards the product respectively; 'Awareness' represents a dummy variable that takes the value 1 for participants who are randomly assigned to the 'Awareness' condition and 0 for participants who are randomly assigned to the 'No Awareness' condition; 'Male' represents a dummy variable that takes the value 1 for male participants and 0 for female participants; 'Age' represents a continuous variable that represents the participant's age; 'Caucasian' is a dummy variable that takes the value 1 for Caucasian participants and 0 for all other participants; 'Education' is a categorical variable with seven categories ranging from "No high school" up to "Graduate degree"; and 'ε' is the random error term.

Estimating the models above for participants with negative opinions about targeted ads showed that awareness about targeting has a significant negative impact on 'LikeProduct' ($\beta_{Awareness}$ = –0.45; p=0.045) and a marginally significant negative impact on 'BuyProduct' ($\beta_{Awareness}$ = –0.40; p=0.08) for these participants. This suggests that awareness about targeting worsens attitudes towards the targeted product and (marginally) reduces purchase intentions towards the targeted product among individuals with negative opinions about targeted ads. Estimating the models for participants who have neutral and positive opinions about targeted ads showed no effect of awareness about targeting on 'LikeProduct' ($\beta_{Awareness}$ = –0.04; p=0.87) and 'BuyProduct' ($\beta_{Awareness}$ = –0.02; p=0.96). This suggests that awareness about targeting does not influence the attitudes and purchase intentions of individuals who have neutral and positive opinions towards targeted ads. The correlation between 'LikeProduct' and 'BuyProduct' is observed to be fairly high (r=0.75, p<0.001) so we created a combined dependent variable by averaging the values of these two dependent variables. This combined dependent variable, 'CombinedProduct', confirmed the results reported above with a significant effect of awareness about targeting ($\beta_{Awareness}$ = –0.42; p=0.048) for the group of individuals who have negative opinions about targeted advertising, and no significant effect of awareness about targeting ($\beta_{Awareness}$ = –0.03; p=0.91) for the group of individuals who have neutral and positive opinions about targeted advertising.

Next, we tested whether the size of the effect of awareness about targeting on attitudes and purchase intentions significantly varies between those who have negative opinions about targeted advertising and those who have neutral and positive opinions. We did this by introducing two additional variables to the basic econometric model above: 1) a dummy variable 'Negative' that takes the value 1 for individuals who have negative opinions about targeted advertising and 0 for those who have neutral and positive opinions about targeted advertising and 2) an interaction term between this dummy variable and the 'Awareness' dummy variable. To ensure that we are comparing the negative group to just the neutral and positive group, we did not include individuals whose consolidated values on our opinions measure fall between 3 and 5. The coefficient on the interaction term is not significant for 'LikeProduct' ($\beta_{Awareness*Negative}$ = –0.38; p=0.25), 'BuyProduct' ($\beta_{Awareness*Negative}$ = –0.39; p=0.26), and 'CombinedProduct' ($\beta_{Awareness*Negative}$ = –0.38; p=0.23). This result suggests that the sizes of the effects of awareness about targeting on attitudes and purchase intentions do not differ significantly between individuals who have negative opinions about targeted advertising ('LikeProducts' Cohen's d = 0.23; 'BuyProducts' Cohen's d = 0.19) and individuals who have neutral and positive opinions ('LikeProducts' Cohen's d = 0.05; 'BuyProducts' Cohen's d = 0.02).

To explore alternative specifications of our moderation hypothesis, we also used a continuous measure of opinions about targeting (instead of splitting participants into groups). This allows us to include all participants in the analysis (even those whose consolidated opinions measure falls between 3 and 5) and treat participants with different values on the consolidated opinions measure differently (as opposed to grouping together everyone with values 3 or below and everyone with values 5 or above). While it is useful to treat the opinions measure as a continuous variable, doing so also introduces additional unexplained variance as we are now including participants who are not sure about their own opinions about targeted advertising (i.e., participants whose consolidated opinions measure falls between 3 and 5). We introduced the consolidated value of our measure for opinions about targeting and the interaction of this measure with our 'Awareness' dummy variable to the basic model described above. The coefficient on the interaction term is not significant for 'LikeProduct' ($\beta_{Awareness*Opinions}$ = 0.09; p=0.27), 'BuyProduct' ($\beta_{Awareness*Opinions}$ = 0.11; p=0.18), and 'CombinedProduct' ($\beta_{Awareness*Opinions}$ = 0.10; p=0.19), suggesting that the effect of awareness about targeting on attitudes and purchase intentions does not significantly vary when opinions about targeted advertising are varied on a continuous 1–7 scale.[3]

Finally, we analyze whether the donation behavior of our participants varied between the 'Awareness' and 'No Awareness' conditions. Participants reported whether or not they would choose to donate to the environmentally friendly charity if they win the lottery, and those who did choose to donate indicated the amount they would like to donate. As the former dependent variable is a dummy variable, we estimate the basic model as a probit. Overall, we find no significant difference in the likelihood to donate between the 'Awareness' and 'No Awareness' conditions ($\beta_{Awareness}$ = 0.03; p=0.79). We also do not find any significant differences in the likelihood to donate between 'Awareness' and 'No Awareness' conditions when specifically looking at individuals who have negative opinions about targeted ads ($\beta_{Awareness}$ = 0.26; p=0.13) and those who have neutral and positive opinions ($\beta_{Awareness}$ = –0.13; p=0.50). We also analyzed the donation amounts using tobit models and find the same null results. Therefore, we do not find evidence that awareness about targeting impacts subsequent donation decisions made by individuals. Results from all models reported in this section are presented in Appendix B.4.

5. DISCUSSION

We investigated the effect of awareness about targeting on individuals' attitudes and purchase intentions towards the advertised product. We find that, among individuals who have negative opinions about targeted ads, awareness about targeting worsens attitudes towards the advertised product and marginally reduces purchase intentions towards the advertised product. We find no effect of awareness on attitudes and purchase intentions towards the advertised product among individuals who have neutral and positive opinions about targeted ads. We also find that 53% of participants in our exploratory study (Exploratory Study) and 33% in our final study (Evaluation Study) reported having negative opinions about targeted advertising. Surveys conducted by previous researchers also suggest that a sizeable proportion of individuals have negative feelings about targeted advertising practices (66% of participants in [39] and 55% of participants in [29] said they do not want targeted advertisements). In addition to uncovering how awareness about targeting impacts attitudes and purchase intentions, we also built and validated a scale that can reliably measure individuals' opinions about targeted advertising.

5.1 Implications

This research raises the important question of whether mere recommendations from the FTC to a self-regulated advertising industry about making consumers aware of targeting are enough to

[3] We conducted the same analysis using the average of only the positive and negative items from our scale (i.e., without the neutral items) and obtained substantively similar results.

protect consumers' privacy. We find that awareness about targeting worsens attitudes towards the advertised product. Since a substantially large proportion of individuals have negative opinions about targeted advertising, our results suggest that the advertising industry is not incentivized to make consumers aware of targeted advertising, as such awareness could lead to lower effectiveness of advertisements. Because targeting practices such as interest-based advertising are not transparent, consumers may be unaware of how their information is being used to influence their purchase behaviors. We believe policy makers should consider introducing and enforcing regulations that require companies to make consumers aware of targeting practices. Another implication of our work is the ability for future researchers to capture attitudes towards interest-based targeted advertising in a reliable manner with our short 6-item scale.

5.2 Limitations

This work presents some limitations. First, we recruited participants from a single participant pool, Amazon Mechanical Turk, in all our studies. It is important to validate our results about targeted advertising with other participant pools. Although previous researchers have shown that Mechanical Turk workers are more demographically diverse than the typical convenience samples of American college students and that results using MTurk samples are similar to more traditional population pools [3, 7], this participant pool is likely to be more savvy about computers than the typical U.S. resident. A second limitation is that our scale validation and evaluation of awareness about targeting on attitudes is only conducted for the interest-based targeted advertising scenario. It is important to validate (and, if needed, modify) our findings for other types of targeting practices such as contextual and remarketed advertising. A third limitation of our work is highlighted by the difference in percentage of negative opinion participants between the Exploratory Study (53%) and the Evaluation Study (33%). This is likely due to differences in how we measured opinions in the Exploratory Study (with open-ended responses) and in the Evaluation Study (with our 6-point scale). A scale is useful in measuring opinions quickly without having to code open-ended responses, but it can introduce bias. A fourth limitation of our work is that we did not collect information about our participants' general offline and online shopping habits, which could have helped reduce statistical noise between conditions. Finally, our Evaluation Study does not account for users' possible overall dislike of ads. It may be interesting to account for that in future work by including a baseline condition with non-contextual, non-behavioral ads.

6. ACKNOWLEDGEMENTS

This research has been in part supported by the National Science Foundation under grants 1012763, 1514192, and 1327992, and by a Google Focused Research Award. We would like to thank all reviewers for their feedback and suggestions.

7. REFERENCES

[1] Alreck, P. L., & Settle, R. B. (1985). *The Survey Research Handbook*, Homewood Ill: Irwin.

[2] Beales, H. (2010). "The Value of Behavioral Targeting. Network Advertising Initiative," http://www.networkadvertising.org/pdfs/Beales_NAI_Study.pdf

[3] Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). "Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk," *Political Analysis, 20(3)*, 351-68.

[4] Bilgic, M., & Mooney, R. J (2005). "Explaining Recommendations: Satisfaction vs. Promotion," *In Beyond Personalization Workshop*, IUI Vol. 5.

[5] Blais, A-R., & Weber, E. U. (2006). "A domain-specific risk-taking (DOSPERT) scale for adult populations," *Judgment and Decision Making*, 1, 33-47.

[6] Brehm, J. W. (1966). *A Theory of Psychological Reactance*, New York: Academic Press.

[7] Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). "Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data?" *Perspectives on Psychological Science*, 6(1), 3-5.

[8] Chen, L., & Pu, P. (2005). "Trust Building in Recommender Agents," *In Proceedings of the Workshop on Web Personalization, Recommender Systems and Intelligent User Interfaces at the 2nd International Conference on E-Business and Telecommunication Networks*.

[9] Clifford, S. (2010). "A Little 'i' to Teach About Online Privacy," *The New York Times*, Jan 26, http://www.nytimes.com/2010/01/27/business/media/27adco.html?_r=0

[10] Cosley, D., Lam, S. K., Albert, I., Konstan, J. A., & Riedl J. (2003). "Is Seeing Believing?: How Recommender System Interfaces Affect Users' Opinions," *In ACM CHI, Vol. 1 of Recommender Systems and Social Computing*, 585-92.

[11] Digital Advertising Alliance (2011). "Your Ad Choices," http://www.youradchoices.com

[12] Edwards, S. M., Li, H., & Lee, J. (2002). "Forced exposure and psychological reactance: antecedents and consequences of the perceived intrusiveness of pop-up ads," *Journal of Advertising, 31(3)*, 83-95.

[13] Egelman, S., & Peer, E. (2015). "Predicting privacy and security attitudes," *ACM SIGCAS Computers and Society*, *45(1)*, 22-28.

[14] Farahat, A., & Bailey, M. C. (2012). "How Effective Is Targeted Advertising?" *In Proceedings of the 21st International Conference on World Wide Web, ACM, Lyon, France*.

[15] Federal Trade Commission (2000). "Fair Information Practices in the Electronic Marketplace," Federal Trade Commission Staff Report, Washington, DC.

[16] Federal Trade Commission (2009). "Self-regulatory Principles for Online Behavioral Advertising," Federal Trade Commission Staff Report, Washington, DC.

[17] Federal Trade Commission (2012). "Protecting Consumer Privacy in an Era of Rapid Change," Federal Trade Commission Staff Report, Washington, DC.

[18] Haws, K., Winterich, K. P., & Reczek, R. W. (2013). "Seeing the world through GREEN-tinted glasses: Green consumption values and responses to environmentally

friendly products," *Journal of Consumer Psychology, 24(3)*, 336-54.

[19] Herlocker, J. L., Konstan, J. A., Borchers, A., & Riedl, J. (1999). "An Algorithmic Framework for Performing Collaborative Filtering," *In Proceedings of the 22nd International Conference on Research and Development in Information Retrieval, New York*, 230-37.

[20] Hill, K. (2012). "How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did," *Forbes*, Feb 12, http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#19b4300034c6

[21] Johnson, H., & Johnson, P. (1993). "Explanation Facilities and Interactive Systems," *In Proceedings of Intelligent User Interfaces*, 159-66.

[22] Leon, P. G., Cranshaw, J., Cranor, L. F., Graves, J., Hastak, M., Ur, B., & Xu, G. (2012a). "What Do Online Behavioral Advertising Privacy Disclosures Communicate to Users?" *In Proceedings of the 2012 ACM Workshop on Privacy in the Electronic Society*, 19-30.

[23] Leon, P. G., Ur, B., Shay, R., Wang, Y., Balebako, R., and Cranor, L. F. (2012b). "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*, 589-98.

[24] Leon, P. G., Ur, B., Wang, Y., Sleeper, M., Balebako, R., Shay, R., Bauer, L., Christodorescu, M., & Cranor, L. F. (2013). "What Matters to Users?: Factors That Affect Users' Willingness to Share Information with Online Advertisers," *In Proceedings of the Ninth Symposium on Usable Privacy and Security,* ACM, 7.

[25] Leon, P. G., Rao, A., Schaub, F., Marsh, A., Cranor, L. F., & Sadeh, N. (2014). "Why People Are (Un)willing to Share Information with Online Advertisers," *In Workshop on Privacy in Electronic Society*, http://www.andrew.cmu.edu/user/pgl/wpes2014oba.pdf

[26] Malheiros, M., Jennett, C., Patel, S., Brostoff, S., & Sasse, M. A. (2012). "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, ACM*, 579-88.

[27] Malhotra, Naresh K., Kim, Sung S., & Agarwal, J. (2004). "Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model," *Information Systems Research, 15(4)*, 336-55.

[28] Mattioli, D. (2012). "On Orbitz, Mac Users Steered to Pricier Hotels," *The Wall Street Journal*, Aug 23, http://www.wsj.com/articles/SB10001424052702304458604577488822667325882

[29] McDonald, A., & Cranor, L. F. (2010)." Beliefs and Behaviors: Internet Users' Understanding of Behavioral Advertising," *In Telecommunications Policy Research Conference, Arlington*.

[30] Melicher, W., Sharif, M., Tan, J., Bauer, L., Christodorescu, M., & Leon, P. G. (2016). "(Do Not) Track Me Sometimes: Users' Contextual Preferences for Web Tracking," *In Proceedings on Privacy Enhancing Technologies, Vol. 2*, 13554.

[31] Morin, C. M., Vallieres, A., & Ivers, H. (2007). "Dysfunctional beliefs and attitudes about sleep (DBAS): validation of a brief version (DBAS-16)," *Sleep*, 30, 1547-54.

[32] Nunnally, J. C. (1978*). Psychometric Theory*, 2nd ed., New York: McGraw-Hill.

[33] Peer, E., Vosgerau, J., & Acquisti, A. (2014). "Reputation as a sufficient condition for data quality on Amazon Mechanical Turk," *Behavior Research Methods, 46(4),* 1023-31.

[34] Purcell, K., Brenner J., & Rainie, L. (2012). "Search Engine Use 2012," PEW Research Center Technical Report, http://www.pewinternet.org/files/old-media/Files/Reports/2012/PIP_Search_Engine_Use_2012.pdf

[35] Rudder, C. (2014). "We Experiment on Human Beings!" OkCupid blog, Jul 28, http://blog.okcupid.com/index.php/we-experiment-on-human-beings/

[36] Scott, S. G., & Bruce, R. A. (1995). "Decision-making style: The development and assessment of a new measure," *Educational and Psychological Measuremen*t, 55(5), 818-31.

[37] Sinha, R., & Swearingen, K. (2002). "The Role of Transparency in Recommender Systems," *In CHI Extended Abstracts on Human Factors in Computing Systems, ACM,* 830-31.

[38] Summers, C. A., Smith, R. W., & Reczek, R. W. (2016). "An Audience of One: Behaviorally Targeted Ads as Implied Social Labels," *Journal of Consumer Research*, ucw012.

[39] Turow, J., King, J., Hoofnagle, C. J., Bleakley, A., & Hennessy, M. (2009). "Americans Reject Tailored Advertising and Three Activities That Enable It," SSRN 1478214.

[40] Valentino-Devries, J., Singer-Vine, J. & Soltani, A. (2012). "Websites Vary Prices, Deals Based on Users' Information," *The Wall Street Journal*, Dec 24, http://www.wsj.com/articles/SB10001424127887323777204578189391813881534

[41] White, T. B., Zahay, D. L., Thorbjørnsen, H., & Shavitt, S. (2007). "Getting too personal: reactance to highly personalized email solicitations," *Marketing Letters, 19 (1)*, 39-50.

APPENDIX

Appendix A

This appendix contains the study materials.

A.1 Scenarios – Exploratory Study

A.1.1 Contextual Targeting Condition

Imagine that you are reading an article on the New York Times website about how to protect your shoes from being damaged by rain and snow. You notice that the advertisement next to the article is for a pair of shoes.

You realize that the advertisement has been specifically targeted to you based on the article that you're reading.

A.1.2 Interest-Based Targeting Condition

Imagine that you are looking to buy a new pair of shoes online. You look at a few different websites that sell shoes. You have not made your decision about which pair of shoes you want to purchase.

A few days later, you are reading an article on the New York Times website. You notice that the advertisement next to the article is for a pair of shoes. You had not seen this particular pair of shoes when you were browsing for shoes some days before.

You realize that your browsing history had been tracked and that it is being used to display an advertisement that has been specifically targeted to you.

A.1.3 Remarketing Condition

Imagine that you are looking to buy a new pair of shoes online. You look into a few different websites that sell shoes. You find a pair of shoes that you like, on a website that you've never visited before. You spend some time looking at this particular pair of shoes and then move on to doing something else. You haven't decided whether or not you want to purchase this particular pair of shoes.

A few days later, you are reading an article on the New York Times website. You notice that the advertisement next to the article is for the same pair of shoes that you were looking at the other day.

You realize that your browsing history had been tracked and that it was being used to display an advertisement that has been specifically targeted to you.

A.2 Self-developed Measures – Exploratory Study

A.2.1 Perceived Control

To what extent do you feel that you have control over this type of targeted advertising? In other words, to what extent do you feel that you can stop receiving this particular type of targeted ads, if you no longer want them?

Please answer this question with respect to the type of targeting described in the scenario on the previous page. [5 points scale from 'Not at all' to 'Very much']

<If the response is 4 or 5 on the previous question then the following question is shown>

Please tell us how you would control this type of targeted ads. In other words, how would you stop receiving this particular type of targeted ads, if you no longer want them? [Open-ended response]

A.2.2 Previous Experience with Online Privacy Invasions Related to Targeted Ads

Have you personally ever experienced an unpleasant or uncomfortable scenario involving targeted ads?

<If yes> Please describe the unpleasant or uncomfortable scenario that you experienced involving targeted ads.

A.2.3 Previous Experience with Online Privacy Invasions in General

Have you personally ever been the victim of what you felt was an improper invasion of your online privacy (irrespective of whether it involved targeted ads or not)? [Modified from Surveys by Louis Harris and Associates and Harris Interactive[4]]

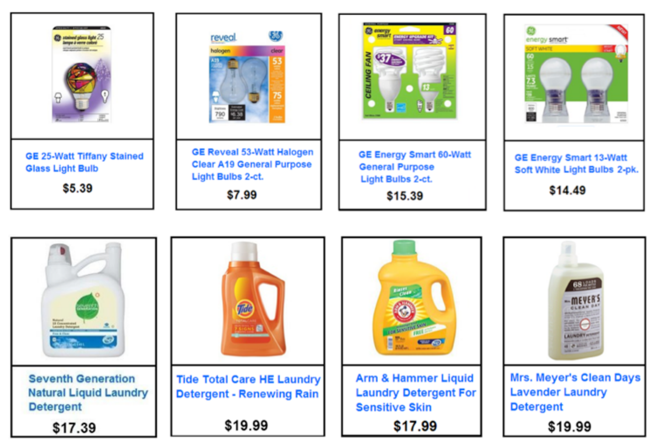
<If yes> Please describe the online privacy invasion that you experienced.

A.2.4 Current Interest in Purchasing Shoes

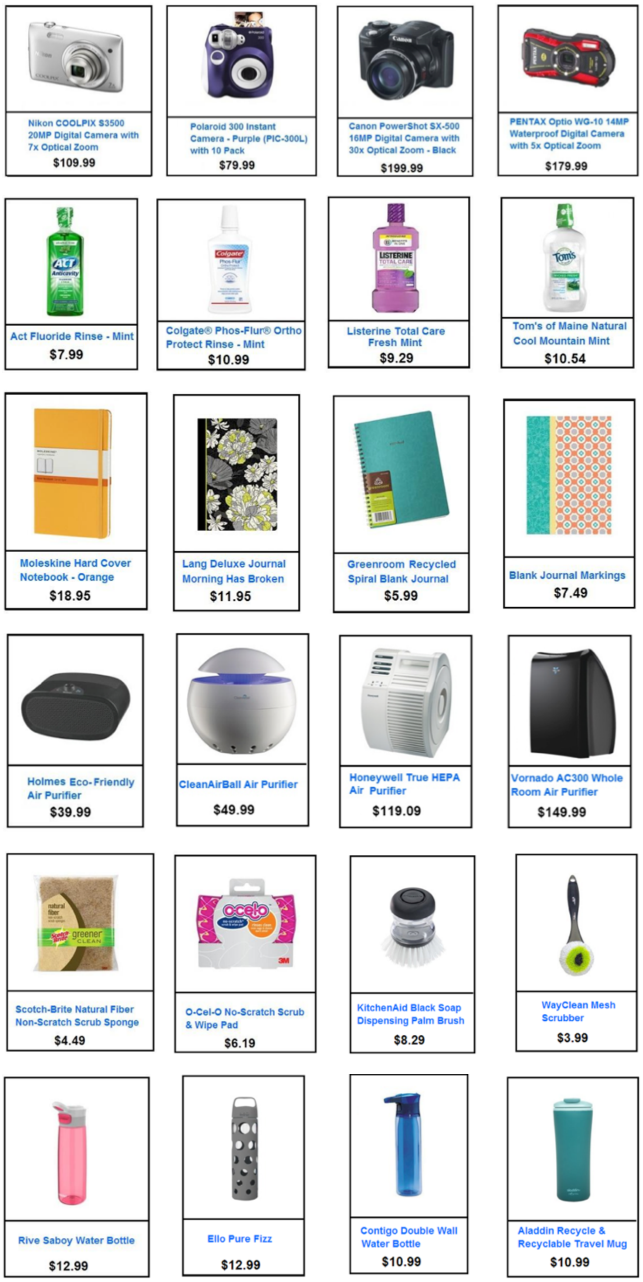
How interested are you currently in purchasing a new pair of shoes, either online or offline? [5 points scale from 'Not at all' to 'Very much']

A.3 Products – Evaluation Study [38]

All 32 products shown below were tested in the pilot study. The four digital cameras and four mouthwashes were dropped from the list of products used in the shopping task in the second phase.

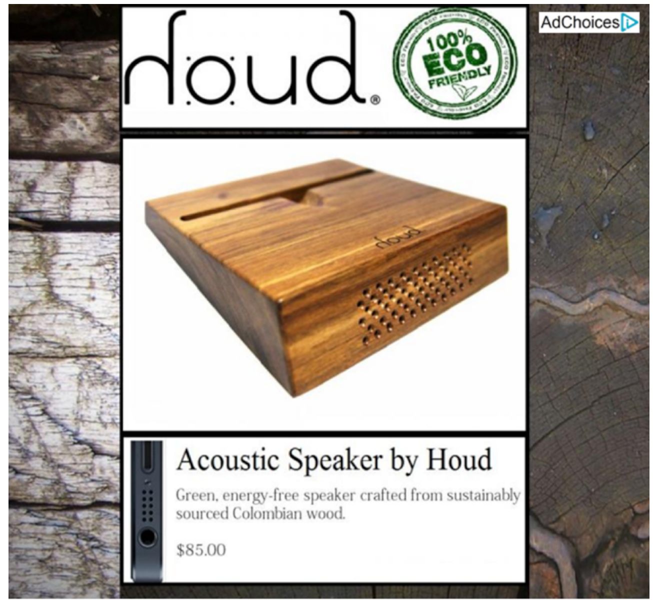


[4] Surveys by Louis Harris and Associates for Southern New England Telephone, September 1-11, 1983, and by Harris Interactive for Business Week, March 3-6, 2000, reported about by The Roper Center for Public Opinion Research: http://www.ropercenter.cornell.edu/public-perspective/ppscan/116/116012.pdf



A.4 Advertisement– Evaluation Study [38]

A.4.1 Awareness Condition



A.4.2 No Awareness Condition

Appendix B

This appendix contains the study results.

B.1 Logistic Regression – Exploratory Study

Odds Ratio:

	(1) Negative Opinions All conditions	(2) Negative Opinions Contextual targeting conditions	(3) Negative Opinions Interest-based targeting and Remarketing conditions
Perceived Control	0.472***	0.567	0.304***
	(0.103)	(0.229)	(0.114)
Previous Unpleasant Targeted Ad Experience	4.040**	7.708*	2.402
	(2.642)	(8.17)	(2.215)
Male	0.662	1.014	0.419
	(0.320)	(0.843)	(0.296)
Age	0.958	0.976	0.937
	(0.025)	(0.042)	(0.038)
Caucasian	0.370*	0.379	0.181*
	(0.212)	(0.327)	(0.171)
Unemployed	0.635	0.719	0.312
	(0.420)	(0.880)	(0.305)
Experience in IT	0.923	0.705	1.30
	(0.576)	(0.770)	(1.26)

*p<0.10; **p<0.05; ***p<0.01; Standard errors in brackets

B.2 OLS Regression – Validation Study A

	(1) Consolidated value of our 6-item scale
Perceived Control	0.140**
	(0.055)
Previous Unpleasant Targeted Ad Experience	–1.163***
	(0.167)
Male	0.058
	(0.168)
Age	–0.013*
	(0.008)
Caucasian	–0.328*
	(0.196)
Unemployed	0.216
	(0.202)
Experience in IT	–0.398*
	(0.219)
Constant	5.222***
	(0.543)

*p<0.10; **p<0.05; ***p<0.01; Standard errors in brackets

B.3 Results – Evaluation Study (Pilot)

Product Category	Product	Mean	Std Dev	T test against scale midpoint 4	Environ-mentally friendly?	Summers et al. (2016) Mean, N = 45	Summers et al. (2016) Std Dev, N = 45
Light bulbs	GE Tiffany Stained Glass	2.63	1.47	$t(143) = -11.20$, $p < 0.0001$	No	2.74	1.42
	GE Reveal Halogen	3.21	1.57	$t(143) = -6.06$, $p < 0.0001$	No	3.47	1.35
	GE Energy Smart	5.37	1.45	$t(143) = 11.31$, $p < 0.0001$	Yes	5.28	1.41
	GE Energy Smart-Soft White	5.29	1.45	$t(143) = -10.67$, $p < 0.0001$	Yes	5.55	1.19
Laundry detergents	Seventh Generation Natural Liquid	5.42	1.34	$t(143) = 12.73$, $p < 0.0001$	Yes	5.36	1.17
	Tide Total Care	3.10	1.53	$t(143) = -7.02$, $p < 0.0001$	No	3.48	1.31
	Arm & Hammer Sensitive Skin	3.26	1.39	$t(143) = -6.34$, $p < 0.0001$	No	3.26	1.12
	Mrs. Meyer's Clean Lavender	4.12	1.73	$t(143) = 1.16$, $p = 0.25$	Can't tell	3.04	1.84
Digital cameras	Nikon COOLPIX S3500	2.85	1.38	$t(143) = -9.94$, $p < 0.0001$	No	2.74	1.48
	Polaroid 300 Instant	2.54	1.34	$t(143) = -13.09$, $p < 0.0001$	No	2.48	1.22
	Canon PowerShot Sx-500	2.71	1.43	$t(143) = -10.82$, $p < 0.0001$	No	2.85	1.62
	PENTAX Optio WG-10	2.93	1.43	$t(143) = -8.96$, $p < 0.0001$	No	3.00	1.74
Mouthwashes	Act Fluoride Rinse	2.97	1.37	$t(143) = -9.09$, $p < 0.0001$	No	3.28	1.46
	Colgate Phos-Flur Ortho Protect Rinse	3.07	1.37	$t(143) = -8.17$, $p < 0.0001$	No	2.89	1.35
	Listerine Total Care	3.10	1.45	$t(143) = -7.50$, $p < 0.0001$	No	3.33	1.38
	Tom's of Maine Natural Cool	4.49	1.56	$t(143) = 3.79$, $p < 0.001$	No	3.98	1.78
Notebook	Moleskin Hard Cover	3.06	1.34	$t(143) = -8.37$, $p < 0.0001$	No	2.78	1.60
	Lang Deluxe Journal	3.15	1.48	$t(143) = -6.92$, $p < 0.0001$	No	2.85	1.71

	Greenroom Recycled Spiral	5.01	1.70	t(143) = 7.17, $p < 0.0001$	Yes	5.07	1.69
	Blank Journal Markings	3.16	1.38	t(143) = – 7.32, $p < 0.0001$	No	2.98	1.73
Air purifiers	Holmes Eco-Friendly	5.00	1.45	t(143) = 8.29, $p < 0.0001$	Yes	4.39	1.71
	CleanAirBall	4.15	1.55	t(143) = 1.13, $p = 0.26$	Can't tell	4.48	1.41
	Honeywell True HEPA	3.89	1.54	t(143) = – 0.97, $p = 0.33$	Can't tell	4.13	1.47
	Vornado AC300 Whole Room	3.92	1.43	t(143) = – 0.70, $p = 0.48$	Can't tell	3.96	1.58
Dish scrubbers	Scotch-Brite Natural Fiber	5.51	1.30	t(143) = 13.97, $p < 0.0001$	Yes	5.07	1.39
	O-Cel-O No Scratch	3.17	1.32	t(143) = – 7.56, $p < 0.0001$	No	3.07	1.44
	KitchenAid Soap Dispensing Palm Brush	3.26	1.34	t(143) = – 6.60, $p < 0.0001$	No	3.48	1.41
	WayClean Mesh	3.33	1.28	t(143) = – 6.25, $p < 0.0001$	No	3.39	1.37
Water bottle	Rive Saboy	4.01	1.63	t(143) = 0.05, $p = 0.96$	Can't tell	4.04	1.93
	Ello Pure Fizz	4.15	1.57	t(143) = 1.12, $p = 0.27$	Can't tell	4.17	1.77
	Contigo Double Wall	3.91	1.62	t(143) = – 0.67, $p = 0.50$	Can't tell	4.07	1.82
	Aladdin Recycle Travel	5.69	1.26	t(143) = 16.08, $p < 0.0001$	Yes	5.39	1.76

B.4 Results – Evaluation Study (Phase 2)

OLS regression coefficients among all participants (N=697):

	(1) LikeProduct	(2) BuyProduct	(3) CombinedProduct
Awareness	–0.084	0.042	–0.021
	(0.120)	(0.129)	(0.117)
Male	–0.128	–0.160	–0.144
	(0.121)	(0.130)	(0.117)
Age	0.003	0.003	0.003
	(0.005)	(0.006)	(0.005)
Caucasian	0.243*	0.220	0.231
	(0.147)	(0.158)	(0.143)
Education	–0.108**	–0.126**	–0.117**
	(0.047)	(0.050)	(0.045)
Constant	4.989***	3.92***	4.456***
	(0.330)	(0.354)	(0.320)

*p<0.10; **p<0.05; ***p<0.01; Standard errors in brackets

OLS regression coefficients among participants who have negative opinions about targeted ads (N=228):

	(1) LikeProduct	(2) BuyProduct	(3) CombinedProduct
Awareness	–0.448**	–0.401*	–0.424**
	(0.222)	(0.229)	(0.213)
Male	–0.048	0.056	0.004
	(0.230)	(0.236)	(0.220)
Age	0.025**	0.023**	0.024***
	(0.010)	(0.010)	(0.009)
Caucasian	–0.184	–0.042	–0.113
	(0.269)	(0.276)	(0.258)
Education	–0.173**	–0.209**	–0.191**
	(0.087)	(0.090)	(0.084)
Constant	4.807***	3.697***	4.252***
	(0.626)	(0.644)	(0.601)

*p<0.10; **p<0.05; ***p<0.01; Standard errors in brackets

OLS regression coefficients among participants who have neutral and positive opinions about targeted ads (N=176):

	(1) LikeProduct	(2) BuyProduct	(3) CombinedProduct
Awareness	–0.042	–0.015	–0.029
	(0.243)	(0.266)	(0.242)
Male	–0.239	–0.477*	–0.358
	(0.244)	(0.266)	(0.243)
Age	–0.010	–0.0004	–0.005
	(0.010)	(0.011)	(0.010)
Caucasian	0.231	0.357	0.294
	(0.337)	(0.369)	(0.336)
Education	–0.061	–0.036	–0.048
	(0.092)	(0.100)	(0.092)
Constant	5.422***	3.828***	4.625***
	(0.659)	(0.720)	(0.656)

*p<0.10; **p<0.05; ***p<0.01; Standard errors in brackets

OLS regression coefficients for models including the interaction between 'Awareness' dummy and 'Negative' dummy among participants who have negative opinions or neutral and positive opinions (N=404):

	(1) LikeProduct	(2) BuyProduct	(3) CombinedProduct
Awareness	–0.034	0.007	–0.014
	(0.248)	(0.262)	(0.242)
Male	–0.125	–0.164	–0.144
	(0.168)	(0.177)	(0.163)
Age	0.009	0.013*	0.011
	(0.007)	(0.008)	(0.007)
Caucasian	0.005	0.121	0.063
	(0.210)	(0.221)	(0.204)
Education	–0.120*	–0.136**	–0.128**
	(0.063)	(0.067)	(0.062)
Negative	–0.133	–0.171	–0.152
	(0.236)	(0.249)	(0.230)
Awareness*Negative	–0.378	–0.391	–0.384
	(0.329)	(0.347)	(0.320)
Constant	5.130***	3.889***	4.510***
	(0.465)	(0.490)	(0.453)

*p<0.10; **p<0.05; ***p<0.01; Standard errors in brackets

OLS regression coefficients for models including the interaction between 'Awareness' dummy and continuous measure for opinions about targeted advertising (N=697):

	(1) LikeProduct	(2) BuyProduct	(3) CombinedProduct
Awareness	–0.419	–0.394	–0.407
	(0.324)	(0.347)	(0.313)
Male	–0.115	–0.146	–0.131
	(0.121)	(0.129)	(0.117)
Age	0.004	0.004	0.004
	(0.005)	(0.006)	(0.005)
Caucasian	0.223	0.196	0.209
	(0.147)	(0.157)	(0.142)
Education	–0.107**	–0.127**	–0.117**
	(0.047)	(0.050)	(0.045)
Opinions	0.052	0.054	0.053
	(0.057)	(0.061)	(0. 055)
Awareness*Opinions	0.087	0.113	0.100
	(0.079)	(0.085)	(0.077)
Constant	4.780***	3.708***	4.244***
	(0.403)	(0.432)	(0.390)

*p<0.10; **p<0.05; ***p<0.01; Standard errors in brackets

Probit results for likelihood to donate to the charity:

	(1) Donate All participants (N=697)	(2) Donate Only 'Negative' group (N=228)	(3) Donate Only 'Neutral & Positive' group (N=176)
Awareness	0.010	0.102	–0.053
	(0.038)	(0.067)	(0.078)
Male	–0.047	–0.108	–0.065
	(0.038)	(0.069)	(0.078)
Age	0.004**	–0.005	0.008**
	(0.002)	(0.003)	(0.004)
Caucasian	0.029	–0.033	–0.009
	(0.047)	(0.081)	(0.110)
Education	–0.019	–0.009	–0.042
	(0.015)	(0.026)	(0.030)

*p<0.10; **p<0.05; ***p<0.01; Standard errors in brackets

Tobit results for amount of donations made to the charity (left censored at 0):

	(1) Donation Amount All participants (N=697)	(2) Donation Amount Only 'Negative' group (N=228)	(3) Donation Amount Only 'Neutral & Positive' group (N=176)
Awareness	–0.077	0.848	–0.858
	(0.405)	(0.694)	(0.750)
Male	–0.488	–0.667	–0.558
	(0.407)	(0.716)	(0.754)
Age	0.036**	–0.040	0.057*
	(0.017)	(0.030)	(0.031)
Caucasian	0.455	–0.424	0.561
	(0.502)	(0.832)	(1.047)
Education	–0.125	–0.246	–0.313
	(0.157)	(0.272)	(0.284)
Constant	–0.383	3.413*	0.574
	(1.112)	(1.969)	(2.028)

*p<0.10; **p<0.05; ***p<0.01; Standard errors in brackets

Using chatbots against voice spam: Analyzing Lenny's effectiveness

Merve Sahin
EURECOM, Monaco Digital Security Agency
Sophia Antipolis, France
merve.sahin@eurecom.fr

Marc Relieu
I3-SES, CNRS, Télécom ParisTech
Sophia Antipolis, France
marc.relieu@telecom-paristech.fr

Aurélien Francillon
EURECOM
Sophia Antipolis, France
aurelien.francillon@eurecom.fr

ABSTRACT
A new countermeasure recently appeared to fight back against unwanted phone calls (such as, telemarketing, survey or scam calls), which consists in connecting back the telemarketer with a phone bot ("robocallee") which mimics a real persona. *Lenny* is such a bot (a computer program) which plays a set of pre-recorded voice messages to interact with the spammers. Although not based on any sophisticated artificial intelligence, Lenny is surprisingly effective in keeping the conversation going for tens of minutes. Moreover, it is clearly recognized as a bot in only 5% of the calls recorded in our dataset. In this paper, we try to understand why Lenny is so successful in dealing with spam calls. To this end, we analyze the recorded conversations of Lenny with various types of spammers. Among 487 publicly available call recordings, we select 200 calls and transcribe them using a commercial service. With this dataset, we first explore the spam ecosystem captured by this chatbot, presenting several statistics on Lenny's interaction with spammers. Then, we use conversation analysis to understand how Lenny is adjusted with the sequential context of such spam calls, keeping a natural flow of conversation. Finally, we discuss a range of research and design issues to gain a better understanding of chatbot conversations and to improve their efficiency.

1. INTRODUCTION
Unwanted phone calls have been a major burden on the users of telephony networks. These calls are often not legitimate (e.g., generated without the consent of the callee) and can be very disturbing for users as they require immediate attention. In the USA, the Federal Trade Commission (FTC) has received over 5 million complaints about such unwanted or fraudulent calls in 2016 [31]. Moreover in 2015, 75% of generic fraud-related complaints reported telephone as the initial method of contact, which raised from 20% in 2010 [29].

The interconnection of IP and telephony networks facilitates voice spam, as it significantly reduces the cost of calls. Voice spam can be performed in many ways, but a common way is to use an auto-dialer equipment to generate vast number of calls to a given (or randomly chosen) list of phone numbers. Once a call is answered, either a pre-recorded message is played (which is called a *robocall*), or the callee is assigned to a live human agent for further interaction. More intelligent auto-dialer equipment (e.g., predictive dialers) can increase efficiency of call-agent scheduling and also check if the call is answered by a person or an answering machine (such as voicemail) [12]. The spam campaigns are often performed by call centers that may belong to legitimate companies, as well as illegitimate organizations.

While the robocalls can be very cheap and very easily disseminated, employing call center agents is often a more costly operation. A 1-minute robocall costs around 4 cents per dial[1], whereas servicing a customer at a call center can cost around 50 cents to $1 per minute [13, 70]. It is also common to utilize overseas call centers (e.g., call centers in India or Philippines [45]), to take advantage of cheap labor. Such call centers still cost around 15-20 cents per minute for outgoing calls [16]. On the other hand, interaction with a live human agent is likely to make the spam campaigns more efficient. In fact, among the 5 million complaints received by FTC, 64% were recorded calls (robocalls) [31], which means the remaining 36% involved human agents. Usually, the number of call center agents are much lower than the number of calls that can be generated by the auto-dialer equipments. As a result, human agents may not have time to answer all the connected calls. Thus, human agents become a limiting factor for fraudsters, whereas the actual cost of generating the call is nearly negligible.

Fighting voice spam is challenging for various reasons. Fraudsters may spoof or block the caller identification (caller ID) information, which makes their identification more difficult. Overseas fraudsters make law enforcement even harder. In many countries, regulators offer consumers to register the *do not call lists* to reduce the number of unwanted calls. However, efficiency of these lists are questionable, as the illegitimate parties do not follow these lists anyway. For instance, the do not call registry in the USA still receives millions of complaints [31]. Moreover, a recent survey shows that 82% of participants did not notice a significant decrease in number of calls after registering to the national do not call

[1] http://www.robocent.com/, http://www.robodial.org/instantpricequote/

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

list [24]. In fact, some forms of calls (such as calls from charities and political organizations) may be exempt from abiding by the do not call lists [32]. In addition, suing telemarketers can be time consuming and costly [6]. Even though important progress has been made on identifying and blocking robocalls (such as mobile applications [18,26], call audio analysis technologies [17], government efforts [28,30]), voice spam remains an open problem.

On the other hand, individuals have been developing their own methods to fight these calls. Many videos where the people are teasing with or scamming back telemarketers and other phone scammers can be found online [4]. Moreover, there exists various recommendations on how to annoy telemarketers and waste their time [2,33]. Due to the cost of human labor, wasting time of one telemarketer leads to a waste of money for the call center, and also saves other people from falling victims to voice spam. For telemarketers, time is money, because each new call they make increases their chance to reach another customer and make profit [3, 47]. However, these individual efforts to stall telemarketers require the callee to waste his time talking to the telemarketer as well.

In this paper, we study an automated way of wasting fraudsters' time and resources (while, at the same time, annoy them). This method employs a chatbot which will act like a legitimate callee and interact with the fraudsters. *Lenny*, to the best of our knowledge, was the first chatbot to become popular for this purpose. It consists of a set of pre-recorded sound files that are played in a specific order to engage in a conversation with a phone spammer.

Although there is no indisputable evidence of this chatbot's origins, some information can be found online. Lenny has been reported to be a recording performed for a specific company who wanted to answer telemarketing calls politely [9]. Later, the recordings were modified to suit residential calls [23]. Moreover, Lenny was inspired from *Asty-Crapper* [7], which was an earlier version of such chatbots, but has not found extensive use. Note that Lenny was not recorded by a professional actor; the voice and age patterns were acted (faked) by a person using his own local accent [9].

Lenny is interesting to study, because it is incredibly realistic and is able to trick many people even without any artificial intelligence or speech recognition mechanism involved. We claim that this success relies on the conversational quality of the recordings. In this paper, we will examine how Lenny is able to stall fraudsters (even up to 1 hour [11]) and discuss the effectiveness of such chatbots to fight voice spam.

Currently, Lenny is provided as a free and open service that allows people to transfer their incoming unwanted calls, using a *warm transfer* or call forwarding.

An important aspect of such chatbots is the usability of the call transfer methods from phone user's perspective (we briefly discuss this in Section 6). However, in this paper, we instead treat the chatbot as a human computer interface and we study the usability of the chatbot in the specific, sequential context of spam calls. Because Lenny's turns fit very well into the conversation, despite being scripted recordings, Lenny has good "usability" as a conversation partner in spam calls. The better its usability, the longer time the caller will waste on the phone, consequently damaging the spam campaign and protecting real users.

We rely our analysis on the call recordings that are available at the public Youtube channel [15]. We select 200 videos from this channel (corresponding to 2,000 minutes of calls) and examine the transcriptions of these calls. We also analyze more than 19,000 call data records (including call date, time and duration) collected by this phone system in the last 1.5 years. Our aim is to shed light on various types of spam calls, different strategies employed by spammers, and also to analyze the conversational properties of these calls to better understand the effect and efficiency of this chatbot.

In summary, in this paper, we make the following contributions:

- We make the first study analyzing a chatbot, which also acts like a high interaction honeypot, to fight voice spam. We observe the different types of spam calls, and evaluate spammers' strategies and interactions with this chatbot.
- We explore the reasons behind the success of this chatbot from an applied conversation analysis perspective.
- Finally, we discuss the challenges in the widespread use of such chatbots and a series of research and design issues.

2. BACKGROUND AND RELATED WORK

In this section we review related work, first on voice spam, then on chatbots and finally on conversation analysis.

2.1 Voice Spam

Voice spam can take many forms and it has been widely studied in the literature. Some studies aim to explore the telephone spam landscape, and better understand the spammers' techniques. A technique frequently used for this purpose is telephony honeypots. A telephony honeypot is a set of phone numbers used to receive spam calls which are received by an automated system (e.g., a VoIP PABX such as Asterisk) and can be interactive (responding to the call and interacting with the caller) or low interaction (not responding to the calls) [35]. Gupta et al., uses a telephony honeypot to analyze 1.3 million calls in a low interaction honeypot [36]. [51] analyzes data from another honeypot that receives robocalls and record the incoming audio. By using certain audio features [20], authors shows that it is possible to identify the infrastructure and the distinct actors behind spam campaigns. Authors find that 51% of robocalls were initiated from 38 different infrastructures [50].

Miramirkhani et al. [54] takes a different approach and studies technical support scams. Authors identify websites advertising scam phone numbers and call 60 of these numbers to interact with the real scammers. They also analyze scammer demeanor (finding that they are usually polite) and the social engineering techniques used by scammers (such as showing various warnings to convince a computer is compromised). Another approach studied in [67] is to look at the linguistic properties of IRS scam calls posted online. This study aims to understand how forensic linguistics may help in the identification of social engineering attempts.

Tu et al. [69] surveys the existing unwanted call prevention techniques and presents an evaluation criteria to assess these. In fact [69] shows that none of the techniques are perfect. While use of chatbots may not be counted as a real spam prevention method, it might be useful to reduce unwanted calls, as it would damage the financial benefits of spammers [61].

2.2 Chatbots

Bots have been built as personas (an artificial but realistic identity) who produce a recognizable type of conduct from the members of such categories (for ex. an "old guy"). Since ELIZA [73], chat bots associate a recognizable identity with a specific ability to produce some linguistic contribution (for instance, turns at talk).

Today, advanced artificial intelligence techniques enable intelligent chatbots, used as personal assistants on smartphones (e.g., Cortana, Siri), application communications (e.g., banking [19]), even as a friend [71]. There are industry efforts to build better and more intelligent chatbots [1,49]. While such advanced chatbots are generally not publicly available, they often have a synthetic voice which is distinguishable from a real human voice. However, it can be expected that such chatbots will keep on improving.

Lenny is not the only chatbot used to fight telemarketers. For example JollyRoger [14] is another similar, but paid, service that hosts multiple chatbots with different personas. However, to the best of our knowledge, Lenny was the first freely available chatbot with a significantly large and public dataset.

2.3 Background on Conversation Analysis

Conversation Analysis (CA) is a sociological perspective which aims at studying the organization of natural talk in interactional order to uncover the seen but unnoticed [34] methodical apparatus which speakers and recipients use in order to solve the basic organizational issues they deal with while talking. Trying to show how the participants to a conversational exchange orient themselves on those methods, CA adopts a descriptive stance, deeply rooted into the detailed analysis of recorded conversational exchanges. Four main apparatus have been isolated and explained, which correspond to four major organizational problems that speakers have to solve.

The first range of issues comes from the management of speakership and hearership between the participants to a conversational exchange. In their famous paper, Sacks et al. [59], present the *turn-taking apparatus*, a model of the methods used to minimize gaps and overlaps while distributing turns in conversation.

In a second classic paper [65], the authors isolate a second pervasive conversational practical problem that speakers tend to solve: *the trouble management issue*. This second apparatus provides a model to explain how speakers repair any trouble in hearing, understanding, or speaking.

The third apparatus deals with the *sequential organizations of actions in talk exchanges*, which we will be using along this paper and therefore deserves a more detailed presentation. Conversationalists assemble their turns in sequences of action which go together. A sequence is an "ordered series of turns through which participants accomplish and coordinate an interactional activity" [53]. A common type of sequence, composed with two interrelated turns has been called an adjacency pair [60, 62, 64]. Question → answer, greetings exchanges, offers → acceptance/rejection or request → acceptance/rejection share many properties of adjacency pairs. Indeed, they consist of two utterances, a first part and a second part (the order), produced by different speakers with an adjacent positioning (contiguous) [60]. The first and second parts fit into specific types, for example, question and answer, or greeting and greeting. The form and content of the second part depends on the type of the first part. Given that a speaker has produced a first part, the second part is relevant and expected as the next utterance. Adjacency pairs share a normative property: Once a first pair part is uttered it becomes conditionally relevant that the other participant should produce the relevant second pair part. In other words, adjacency pairs point to the normative expectations that are embedded into the ways we order turns at talk as pairs.

The fourth apparatus aims at *clarifying how speakers use membership categories during talk exchanges*. [57] and [58] discussed how conversationalists use categories to recognize, identify, describe or infer about people. This range of topics have been explored in a sub CA area called "Membership Categorization Analysis" (MCA [42,44]). Identities, such as "elderly", can be displayed within and through the sequential organization of talk, without being explicitly referred to. Most CA studies have demonstrated how categories and identities are made demonstrably relevant by the participants themselves in the detail of their talk.

3. DATA COLLECTION & METHODOLOGY

Telephony honeypots commonly use large sets of unused phone numbers, such as new (previously not allocated) phone numbers or, better, numbers which have been returned by users who receive too much spam. Such phone numbers are then directed to an IP-PBX (IP based Private Branch Exchange). An IP-PBX uses a set of phone lines to receive calls and allows to process (e.g., answer, record, forward) these calls. Low interaction honeypots will let ring the call or hangup and record the call metadata. In addition to this, a high interaction honeypot will answer the call and interact with the caller. A difficulty for setting up high interaction honeypots is that in many countries recording the call requires both caller and callee agreement, otherwise, recording without agreement could be considered as illegal wiretapping. Asking for permission would however change caller behavior or raise suspicion. Indeed, as it is uncommon for callees to request permission to record this would bias the study. The recordings we used in this paper were all conducted in a country and under conditions which make those recordings legal.[2]

3.1 Lenny's Interactive Voice Response (IVR) System

Lenny's voice recording are publicly available, and our study focuses on one particular deployment which made audio recordings available and attracted a significant amount of interest [8, 22].

In Lenny's particular implementation (Figure 1), incoming phone calls are answered and the set of audio recordings are

[2]We omit details to preserve the anonymity of the PBX maintainer.

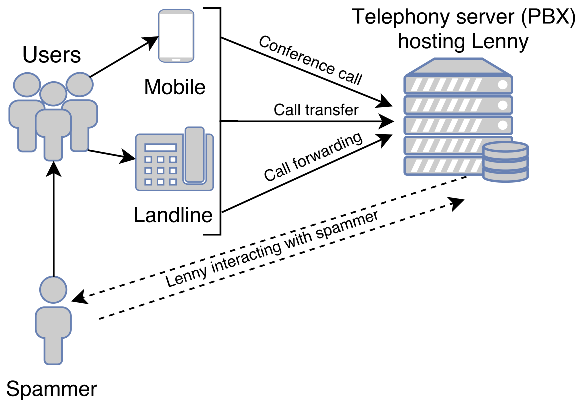


Figure 1: Deployment setup and usage.

played one after another, to interact with the caller. There is no speech recognition or artificial intelligence to select or modify Lenny's answers, the same set of prompts is always used in the same order. This is controlled by an Interactive Voice Response (IVR) script which allows simple scripting and detection of silences.

The script starts with a simple "Hello, this is Lenny." and will wait for the caller to take his turn. If he does not respond within 7 seconds, the server switches to a set of "Hello?" playbacks until the caller takes his turn. However, if the caller speaks, the IVR script waits until he finishes his turn. The script detects the end of the caller's turn by detecting a 1.55 second long silence period, at this point it will play the next recording. When the 16 distinct turns that are available have been played, it returns to the 5th turn (the 4 first prompts are supposed to be introductory adjacency pairs) and continues playing those 12 turns sequentially, forever.

The PBX server hosting Lenny is reachable both via a SIP URI and via a landline number. Some common methods to transfer a call to Lenny are (Figure 1):

- When a phone user identifies a spam call, he asks the spammer to hold on for a second, then either transfers the call to the phone number of the PBX server or creates a 3-way conference call, and lets Lenny interact with the spammer.[3] In this case, the caller ID logged on the PBX server will belong to the phone user.
- A user can directly forward previously known (blacklisted) spam numbers to Lenny. In this case, Lenny will be the first respondent of the call, and the PBX server will log the spammer's caller ID.

It is estimated that around 500 users are using this service, as the calls are targeted to real users they sometimes contain private data, such private data is curated before the calls are made public.

3.2 Public Dataset and Selection

We use data collected by a popular deployment of Lenny for which a set of call recordings are available online on Youtube [5]. As of November 14th, 2016, the Youtube channel contains 487 unsolicited calls answered by Lenny, with an average call duration of 09:43 minutes. In addition to this, we obtained the PBX server call logs (call date, time and duration) for 19,402 spam calls sent to Lenny over 18 months (from 06/17/2015 to 12/17/2017).

Among the 487 public call logs, we select 200 calls randomly, but preserving the call durations distribution (Figure 2). We also include some interesting outliers, like a 1-hour call.

We then used a commercial transcription service to facilitate the analysis of the call recordings. Over 2000 minutes of Lenny calls were transcribed with verbatim transcription and timing of each turn of the conversation. We chose a professional transcription service over a speech recognition tool (like in [51]) in order to obtain the high transcription accuracy required for conversation analysis. Finally, we converted selected fragments of transcripts to the Jeffersonian transcription notation [46] required for very fine grained analysis.

3.3 Limitations of the Dataset

While this dataset is relatively large and instructive on the discussions between abusive telemarketers and Lenny, it comes with a few limitations.

First, the audio recordings publicly available on Youtube were selected by the owner of the PBX server subjectively, with a changing criteria over 3 years.

Second, the call recordings are not always complete, they only contain the part of the call that is handled by Lenny (after it has been transferred) and some parts have been edited to remove personal information.

Finally, the IP-PBX does not always receive the caller ID information of the spammer, but the caller ID of the user transferring the call. As a result, it is not possible to precisely know the spammers' caller IDs and to use this in our analysis. Moreover, a user may arbitrarily transfer only a subset of the spam calls he receives, so the coverage is limited compared to the other honeypots which do not require a human to transfer the call [36, 51].

Nevertheless, this dataset is very interesting to understand and analyze the audio conversations between a telemarketer and an automated system.

4. ANALYZING THE SPAM LANDSCAPE

In this section we will analyze the voice spam landscape, comparing our observations with previous work. We will also analyze how call agents behave and how their behavior vary according to the type of the spam call.

4.1 Observations on Call Logs

We observe several trends on the spam calls, using the 18-months dataset of 19,402 calls. Figure 3 shows how the calls are distributed over the days of a week and hours of a day. Majority of the calls were made on weekdays and business hours, which is in line with the findings in [36].

Figure 4 shows the distribution of the call durations (in minutes). In particular, 78% of the calls were less than 2-minutes long. On closer inspection, many of those short duration calls were due to call forwarding problems. In other, more frequent cases "abandoned" calls were dialed by a pre-

[3] In a conference call, the user can mute his phone and does not need to interact.

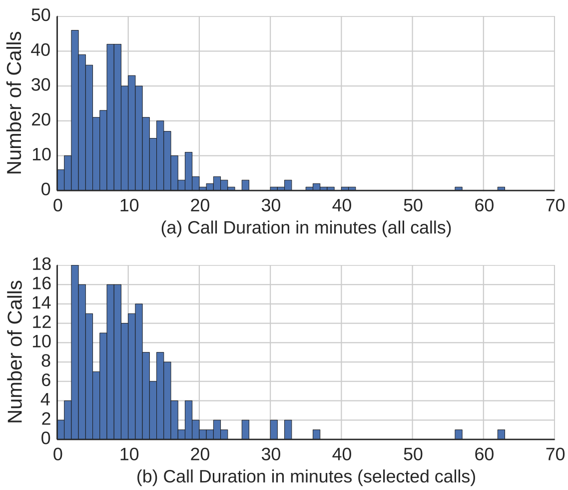


Figure 2: Histogram of call durations uploaded on Youtube channel, (a) all calls as of November 14th, (b) selected calls.

dictive dialer, but were not transferred to a human agent afterwards, or dropped by the caller. Unfortunately, we did not have access to all recordings of such calls and we therefore do not have detailed measurements on this aspect. We assume that the calls longer than 2 minutes contain real conversations of spammers with Lenny. Considering the 4094 calls that are longer than 2 minutes, we find that Lenny stalled spammers for more than 385 hours in 18 months, with an average call duration of 5.6 minutes.

Due to privacy concerns, the PBX logs we obtain do not contain any caller IDs. Moreover, as explained in Section 3.1, caller IDs received by the PBX may belong to the spammers, and may be spoofed. Therefore, we cannot present statistics on the increase or decrease of spam calls experienced by individual users over time. However, we present the monthly distribution of calls in Figure 5. Note that the increase in calls may result from the increase in the popularity of the PBX server among the online community.

4.2 Analysis of Call Recordings

Transcriptions of call recordings provide valuable insights on different types of unsolicited calls the customers experience, and the strategies frequently used by fraudsters to convince customers.

Initially, we isolate the spammers' turns in each transcript, tokenize the words and use k-means clustering algorithm (with k=15) to cluster the spam calls. Then, we manually examine the results and end up with 22 clusters. Upon further examination, we create a broader classification of spam types: fundraising, telemarketing (targeting home owners, business owners or personal) and scam calls. Table 1 presents the descriptions of different spam calls in each category.

In general, fundraising calls aim to collect donation for political organizations and charities. Telemarketing calls either try to identify potential customers for a business (referred to as 'lead generation' calls in telemarketing terminology [21]) or try to sell a product. On the other hand, scam calls include all sorts of calls trying to deceive people into making

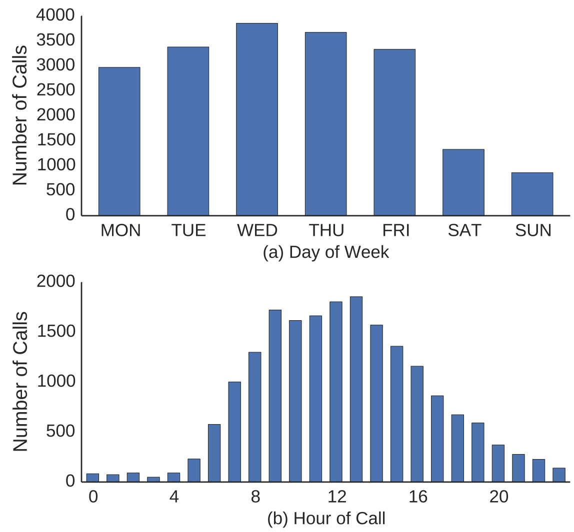


Figure 3: Histogram of calls by (a) days of a week and (b) hours of a day. Note that time zone of callee might be different from time zone of the PBX server in some cases.

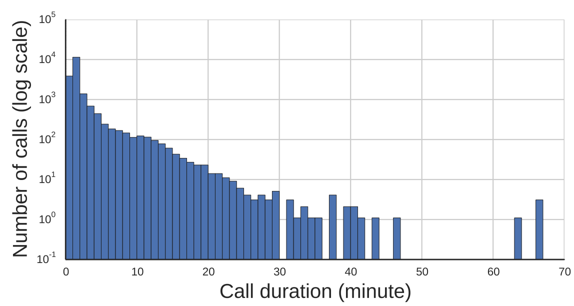


Figure 4: Histogram of call durations covering 18 months.

a payment or revealing sensitive information to gain illegitimate benefits.

We observe that a spam call usually starts with a composition of the following turns from the caller (see [63] for an extensive analysis of informal call beginnings):

- Greeting (e.g., 'Hello')
- Self identification (Name of the call agent)
- Company identification (Name of the business)
- Warm up talk (e.g., 'How are you today?')
- Statement of the reason of the call
- Callee identity check (callee's name and attribute)

While identifying the company, spammers often use phrases assuring the legitimacy of the business. While the telemarketers use phrases like "licensed, bonded, insured company", scammers are likely to use a illegitimate or fake company name referring to a well-known institution (e.g., 'Windows service center' or 'US Grants and Treasury Department').

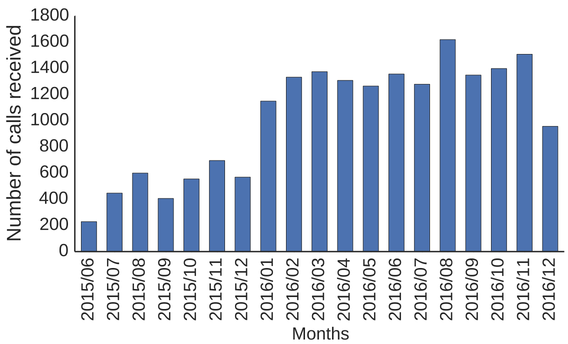


Figure 5: Number of calls received by the PBX server each month.

However, here we do not judge the legitimacy of the involved businesses in telemarketing and fundraising calls. Nevertheless, these calls are unwanted (as the user transferred them to Lenny) and often aimed to manipulate customers.

Callee identity check usually aims to verify that the callee is the 'decision maker' (e.g., the owner of the house or business) or he is in need of a certain opportunity (such as lowering interest rates for credit card debt).

To better convince the customers, spammers make several promises throughout the call, such as they will give a free estimate with no obligation, cancellation is easy or free, the price is all inclusive or there will be a lifetime warranty. Another strategy is to pressure the customer for a quick decision. For example, some scams start by congratulating the person to make him believe that he won something and this is a limited time offer (e.g., "valid only for today"). On the other hand, some calls start with a threatening scenario such as "your computer is getting infected", "your air duct system is badly contaminated" or "there are 8,000 home invasions everyday in the US".

During the call, spammers ask several questions, some of which are summarized in Table 1. We believe that even if the customer does not qualify or does not accept the offer for the moment, this information is collected to broaden and verify information on customers, which can be used for more efficient advertisement in the future [66].

The final purpose of the spammer is often to convince the customer to make a payment (e.g., by giving credit card information or home address for the bill), or to get an appointment for further interaction. We frequently observe that the spammer does not give the customer an option to decline. Instead, he asks to choose between two different products or services. For instance:

- Donation for a political party: spammer asks if the customer wants to donate \$625 or \$500, later in the call \$425 or \$375, and later, \$250 or \$100.
- Appointment for home improvement technician: spammer asks if the customers prefers 2:30pm or 4pm.
- Medical equipment: spammer asks if the customer needs a knee brace or a back brace.

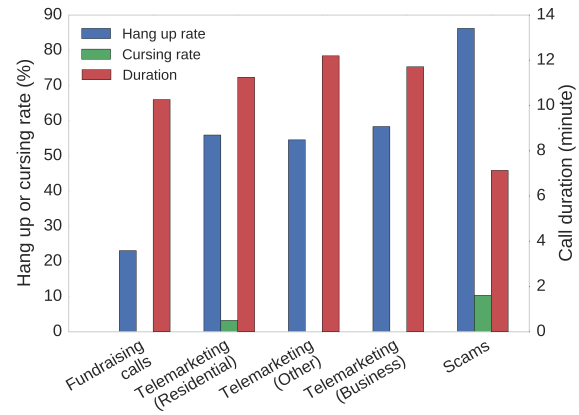


Figure 6: Interaction of different type of spammers with Lenny.

4.2.1 Interaction with Lenny

Before we analyze Lenny's conversational properties, we would like to present some statistics on how spammers interact with Lenny. In our dataset consisting of 200 calls, spammers on average spend 10:13 minutes talking to Lenny. These conversations include an average of approximately 58 turns (an exact calculation is difficult due to overlapping speeches). Moreover, 72% of calls contain Lenny's set of scripts repeated more than once. On average, a caller hears 27 turns of Lenny, which corresponds to 1.7 times repetition of the whole script. These results show that Lenny is a quite successful chatbot in continuing the conversation.

Surprisingly, in only 11 calls (5% of all calls), the caller realizes and states that he is talking to a recording or an automated system. Additionally, 5 of them notice the repetitions in Lenny's turns and state that "something is wrong" with Lenny. 7 spammers think that Lenny has dementia or alzheimer and/or try to contact his nurse, whereas 4 other ones ask Lenny if he is playing a prank on them. 2 of the spammers who realize Lenny is a recording say that they are still getting paid for the call, one even threatens him to be calling every morning at 8:30 am [10]. Moreover, several spammers aggresively try to interrupt Lenny by shouting phrases like "sir please stop" or "listen to me", or even by clapping hands.

In Figure 6, we analyze how spammers' behavior vary in relation to the different type of spam calls. The *hang up rate* shows what ratio of the spammers hang up the call on Lenny, without a proper closing turn. Even though Lenny's neverending turns make it hard to leave the conversation, some spammers try to politely end the conversation by pretending that they are not able to hear Lenny or they have to leave for a meeting, and saying that they will call back at a later time. The *cursing rate* shows the ratio of spammers from each category that use bad language and swear words. Finally, we present the average call duration for each category as well.

Looking at Figure 6, we can say that fundraising calls are more polite than others. Such calls often come from charities and political organizations, who usually care about their reputations and impressions they make. Telemarketing calls

Table 1: Categorization and description of spam types.

Category	*Descriptions of spam types*	*Requested personal information*
Fundraising *(14 calls)*	**Political** calls to collect donations for political parties or organizations **Charity** calls to solicit contributions for charities	- Political affiliation - Credit card information - Email
Telemarketing targeting home owners *(93 calls)*	**Home improvement** calls offering discounts and free price estimates on various work needed around the house, like **window&door replacements** **Furnace and air duct** cleaning/upgrade promotions **Solar energy** calls offering free installation of solar panels to provide lower rates on electricity bills **Security alarm system** companies offering installation of a free (or discounted) alarm system (but requiring a monthly monitoring fee) **Energy providers** offering discounted, flat rate utility bills **Communication providers** offering phone/TV/Internet bundles	- Age of the house - Age of furnace or air conditioning - If the callee is married or single - Recent electricity bill, current energy provider - Recent Internet bill, current provider - TV count in the house - Home address
Telemarketing targeting business owners *(12 calls)*	**Office supply** company offering discounts and free shipping on orders **Business directories** offering premium business listing	- Business name - Location
Other consumer centric telemarketing *(22 calls)*	Medication or medical equipment offers, extended car warranty, newspaper and magazine subscriptions	- Medical history, pain problems - Car mileage - Credit card or check address
Scams *(59 calls)*	**Technical support scams** offer a fake tech support service and request money **Vacation scams** offer a free vacation, but the customer needs to pay for government/port taxes **Credit card scam** offers lower interest rates on credit card debt, but the customer gets no real benefits **Advance-fee and cash advance scams** promise a sum of money, or funding for businesses, but the customer needs to pay up-front fees **SEO scam** offers guaranteed rankings on search engines (claiming relation to a well known company)	- Full name, email - Credit card information - Credit card balance - Current bank interest rate - Business profit - Business name, website

show similar characteristics, regardless of the call target. On the other hand, scammers are the rudest callers with 89% hang up rate and use of swear words in 10% of calls.

We also apply Chi-squared and Fisher's exact tests on hang up rates and observe a statistically significant relation between the hangup rate and spam category (telemarketing, fundraising, scam): for the significance level of 0.05, p-values are less than 0.0001.

As opposed to the polite demeanor observed in [54], we find tech support scammers to be particularly rude against Lenny, with 100% hang up rate and 20% cursing rate, probably because Lenny does not comply with their instructions.

Scam calls also have a noticeably shorter average call duration compared to other spam types. Applying a two sample T-test (p=0.05) for each pair of the three categories shows that the duration of scam calls are indeed significantly different from both fundraising and telemarketing calls. Again, a possible reason is that the scammers do not want to waste time with Lenny, once they realize that Lenny will not answer their questions or do what they ask. However, fundraising and telemarketing calls do not have a significant difference between them.

5. USABILITY OF LENNY AS A CONVERSATION PARTNER: AN APPLIED CA APPROACH

Lenny's efficiency is closely related to how the pre-recorded, pre-defined turns are able to do the job that is inevitably, and unremarkably done during each call, to solve the multiple, interrelated tasks which come from the fourth levels of the organization of talk (Section 2.3).

1. L: hello: eh eh this is lenny!
2. L: hi uh uh ss- sorry, I'b- (0.3) I can barely hear you there?
3. L: °ye-° ye:s yes yes
4. L: oh good >yes, yes, yes<.
5. L: uh yes, yes. uh: uh uh, someone, someone did (.) did ss- say last week (abou-) someone did call last week about the same (.) thing, .h wa- was that was that you?

Figure 7: CA transcript of the first pre-recorded Lenny's "turns" (formatted with [46]).

5.1 The Structure of Lenny's Turns

Figure 7 shows the first five turns of Lenny (T1 to T5). After a direct, informal reception of the call in which he gives his first name (T1), Lenny introduces a hearing issue (T2), then produces a first "yes" turn (T3), followed by a more enthusiastic one (T4), and a last "yes" turn which has a second part, a verification question about a past event (T5). From a CA perspective, those pre-recorded items have both sequential and turn-constructional features, which refer to the organization of sequences in interaction and to the organization of turn management.

Each turn is supposed to play a specific role in the construction of the sequences of actions which will be built in each call. Though it is not possible to say in advance what those sequences will consist of, the design of the turns will foster a specific sequential development. CA stress on two features of turns in sequence: a turn is addressing the immediate preceding talk (it is "context-shaped" [38,60]); and a turn is projecting some next action (it is "context-renewed"). Some turns become parts of a two-unit sequence, the Adjacency Pair. Though T1 and T2 have been designed as first pair parts, which project second pair parts, both T3 and T4 are designed as second pair parts of an adjacency pair (i.e., that they are supposed to follow a question, a proposal, request, etc.). Moreover, the beginning of T4 adds two other components: "oh", a turn-initiated particle.[4] which has been demonstrably analyzed as a change-of-state token [37], a prefaced response [39] and is frequently linked to the making of assessment [40], in particular when it is followed by some assessment token, such as "good", and the three enthusiastic "yes" which end the turn. This design suggests that the turn can get some local sense from several second positions besides the yes/no questions and confirms that this turn is supposed to be "backward looking" [37]. T5 sounds as a verification question and presupposes that the reason for the call has been previously introduced by the caller. Then this turn has a distinctive sequential property: it has been built to occupy a more specific sequential position in the call (the position after the reason for the call).

From the perspective of CA, Lenny's turns share another feature: they have been designed to display repair related features. Almost all turns display self-initiated self-repairs (T.1, 2, 3, 5). The initiations are produced through cut-off (l.&, 2) or "uh" types. The proper repairs are produced through repeats (T5) or transformations (T2). It has been suggested that a high frequency of "disfluencies" in talk features the class of age of the speaker [43]. Along with the pitch of his voice, such disfluencies facilitate the possible recognition of Lenny as an "old man" and bring an easy explanation for some other understanding troubles which might occur. The availability of this membership category [58] can be used by telemarketers, in some calls, as a relevant account for other features of Lenny's talk.

Inspecting Lenny's turns in isolation is not sufficient enough to understand how Lenny can be so efficient in so many different calls. This efficiency is locally built in each call development. Once embedded into a real call, Lenny's turns display an understanding of prior turn and brings new material to be understood by his co-participant. This in situ inspection of Lenny's turn is inevitably made, with more or less care, by the participants, in order to build their own contribution and to fit each new turn into the ongoing conversation. This is what CA calls the "next-turn proof procedure" [59] and what explains the various, flexible ways in which Lenny's turns can play their part in some calls.

5.2 An analytic insight on the opening section of Lenny's turns

On one hand, most telemarketers use very detailed scripts while talking to a prospect. For this reason, the call trajectories might seem to be even more routinely organized than the informal talk on the phone. On the other hand, the Lenny corpus displays different types of calls (See Table 1) and several different caller objectives. This tension between routine and diversity can be seen in the various sections which compose the beginning section and is solved, in some way, by Lenny's style of participation. In the limited space of this paper, we will only examine the beginning section, because it is often a strategic place in which the trajectory of calls is prepared and launched.

In this paper, the beginning section will refer to the talk which has been produced before the production of the reason for the call.

5.2.1 Calls with minimal beginning section

Some calls do not display any beginning section: the reason for the call is given in the first possible position in the call, just after Lenny's first turn.

In a very few calls, this is done without any self identification of the caller (Fragments 1, 2) or with a minimal identification (Fragment 3).

Fragment 1.

```
1.  Lenny: hello: eh eh this is lenny!
2.  Adam: yeah mister lenny, you have been chosen to get a
          lower interest rate, so (I believe) you have
          pressed one to get a lower interest rate right?
3.        (0,6)
4.  Lenny: hi uh uh ss- sorry, I'b- (0,3) I can barely hear you
          there?
5.  Adam: I'm saying so I believe you have pressed one to get
          a lower interest rate, right?
6.        (0,6)
7.  Lenny: °ye-°  ye:s yes yes.
```

In turn 2 of Fragment 1, the caller goes directly to the point, without even a self identification, an identity check question to the callee, a greeting or any other item. The caller addresses the callee with the first name he has given in his first turn. This is the first adaptation of the script to the specificities of this call with Lenny. Then the business of the talk is addressed with no more preparation but it refers to a previous action which the caller has been accomplished on the phone ("pressing one"). Let us remark that true or not, its aim is to focus the attention of the callee to bring an answer in the next turn and to attend to the call. In this sense, this turn makes the organizational job to drive the callee's attention right to the business of the call. From a sequential perspective, this is a not any kind of yes/no question [56]: the "polarity" of the interrogative embodies a preference for a "yes". From the management of turn perspective, an important consequence here is that a positive answer will give the floor back to the caller. Then it provides the caller with a convenient, quick way to get into the call and to project a next slot for his following question. How Lenny's first turns handle those opportunities? First, the T4 initiates a repair sequence, which is answered by a partial repetition of the first caller's turn, the last yes/no question. Because Lenny's next turn is precisely designed as a "yes" answer, it does the job, selects the preferred answer – a "type-conforming response" [56] and the caller can ask the next question.

[4]"oh" is the "second most common turn-initial object in English conversation" [41,55]

Fragment 2.

1. Lenny: hello: eh eh this is lenny!
2. Caller: injured, retired and elderly fop members, as well as create and maintain a state memorial for officers tragically killed in the line of duty. now sir when you receive your pledge kit, can the pennsylvania fraternal order of police foundation count on you for a fully tax deductible donation?
3. (0,6)
4. Lenny: hi uh uh ss- sorry, I'b- (0,3) I can barely hear you there?
5. (1)
6. Caller: oh I was just saying the goal of the drive is to help provide assistance to families in local lodges when fop members are killed in the line of duty. assist injured, retired and elderly fop members, as well as create and maintain a state memorial for officers tragically killed in the line of duty. now sir when you receive your pledge kit, can the pennsylvania fraternal order of police foundation count on you for a fully tax deductible donation?
7. (0,6)
8. Lenny: °ye-° ye:s yes yes.

In the donation call (Fragment 2), the caller again rushes into the presentation of the reason for the call, but in a somewhat different way. Though this long turn is finished with a yes/no question and then orients to a third turn for the caller, the donation proposal has been prefaced by a long attempt to emotionally engage the callee into a supportive action for police officers and their families who are in difficulty. Thus, it aims to trigger a yes answer. Lenny's turn fit very well into this second beginning.

In the two following fragments, the caller rushes into the business of the call after a short self identification.

Fragment 3.

1. Lenny: hello: eh eh this is lenny!
2. (0.6)
3. Lenny: hi uh uh ss- sorry, I'b- (0.3) I can barely hear you there?
4. Jeff: hi, sir, my name is jeff, I'm calling you from business loans center (.) to offer you a cash advance for your business. are you interested?
5. (0,6)
6. Lenny: °ye-° ye:s yes yes.
7. Jeff: oh. (.) okay, let me ask you a few qualifying questions then. (.) are you the business owner?
8. (0,6)
9. Lenny: oh good >yes, yes, yes<.

In Fragment 3, this self identification is completed by an identification of the institution he is calling on behalf. Then the caller brings immediately a question to the attention of the callee. This is a cash advance proposal oriented to business owners. From this, we can guess that the phone number has been found on a list of business firms. In such cases, because the reason for the call has been built as an attention getting device, the identity verification check is made after this turn. Here, after Lenny's first "Yes" answer (T.6), the caller adds an identification question which is formatted as a question about the callee's professional status.

Fragment 4.

1. Lenny: hello: eh eh this is lenny!
2. Caller: oh hi I'm calling at HVAC heating and air we're going to be in the area during the maintenance for thirtynine dollars if you haven't done this yet as a as a promotion we're doing for the next month ok but it does include air duct cleaning and under filter and a safety inspection for the furnace,
3. (0,6)
4. Lenny: hi uh uh ss- sorry, I'b- (0.3) I can barely hear you there?
5. Caller: ho. uh ha- have- have you had a maintenance done on your furnace this winter ?
6. (0,6)
7. Lenny: °ye-° ye:s yes yes.
8. Caller: ok we are offering it for a half price right now.

In Fragment 4, the caller quickly identifies the firms he is calling from, to announce the reason for the call, a promotional offer. Note that after the repair initiation of Lenny (T4 here), the caller does not repeat the promotional offer but recycles it as a verification question which gives him the floor back to re-introduce the offer in the following turn, after Lenny's "Yes" turn (T7 here).

5.2.2 Calls with beginning section: a progressive entry into the business of the call

In most calls, however, the caller does not introduce the reason for the call directly in the first turn. He first greets Lenny back, adds a self identification and/or a "how are you" question.

Fragment 5.

1. Lenny: hello: eh eh this is lenny!
2. Caller: hi sir good afternoon this is michelle with cool duct air-conditioning and heating how are you today ?
3. Lenny: hi uh uh ss- sorry, I'b- (0.3) I can barely hear you there?
4. Caller: ow okay. hu this is michelle with cool ducted air-conditioning and heating. (.) Is that bett[er ?
5. lenny: [°ye-° ye:s yes yes.
6. Caller: okay, uhm : I was calling about the summer maintenance for your central air conditioning,
7. (0,6)
8. Lenny: oh good >yes, yes, yes<.
9. Caller: ok we're running a special right now it is more than fifty percent off and we do a 50-point tune-up on your unit now with our service we replace your filter with a free reusable filter and we talked off the refrigerant up to a pound at no charge ok it's automatically a hundred dollars savings ok now uhm the technicians going to also do a complete mold and mildew inspection flushing vacuum out the drain line to get rid of the build-up and then treat it without the side tablets ok so on it will prevent the build-up of mold and mildew and kill bacteria so you're getting better quality air in your home he'll also check your ducts to make sure there's no tears or separations anywhere that you're losing energy ok we wanted you should save up to twenty percent on your monthly electric bill ok now he'll also check your air handler, calibrate the thermostat make sure it's accurate the voltage and

the amperage on the motors the starting capabilities to make sure the A/C is turning on and off when it's supposed to fill sanitizer evaporation oil and oil and lubricate all the moving parts. ok now more importantly you'll get the state certified report to validate your warranty cap that's your proper documentation. okay and our special right now is only forty-four dollars not a penny more that includes the taxes you're 50 point tune up your free filter free refrigerant along with a hurricane safety inspection also free and there is no trip charge or we could take care of you tomorrow.

In the opening of Fragment 5, the presentation of the reason for the call is prefaced with a first multi-units turn in which the caller introduces a greeting, a self identification, and an identification of the firm she is calling for, before adding a "how are you" question. Note that the next turn, in which Lenny initiates the "hearing" repair, is answered as a partial repeat from which the "how are you" is now absent. Michelle completes this turn as a hearing check, with a yes/no question. Then Lenny's first "Yes" turn fits well in this sequential environment and displays a confirmation. In turn 6, Michelle, the caller, introduces a first characterization of the "reason for the call", which is often briefly presented in the opening section. Lenny's enthusiastic second "Yes" turn (T.8) sounds, in this sequential context, as an authorization to expand the previous announcement.

In fragments 6 and 7, the identification questions have been introduced before the reason for the call. In Fragment 6, the identification question is about the callee's name, while in the next fragment (7), the identification sequence is relative to a role.

Fragment 6.

1. Lenny: hello: eh eh this is lenny!
2. Caller: hello, am I speaking with mister ( )
3. Lenny: hi uh uh ss- sorry, I'b- (0.3) I can barely hear you there?
4. Caller: oh I'm sorry about that, can you hear me now mister ( )
5. Caller: °ye-° ye:s yes yes.
6. Caller: okay, okay, ah mister ( ) my name is lui by the way. how are you doing today?
7. Caller: oh good >yes, yes, yes<.
8. Caller: okay, good. um I was just giving you a call at this time because one of my reps had spoke with somebody in your home. um today I was just following up just to see if anything has changed since we last spoke with you. if you need any free estimates for any home repairs that you're trying to get around to or you may have in mind at all?

The first identity verification question has been asked in turn 2 by the caller. Then the caller produces a hearing check in turn 4, using the name of the prospective callee. Because Lenny's next turn is the first "Yes" turn, it displays an embodied acceptance of the addressee term and then closes the identity issue. After a "how are you" question, who confirms the expected progression of the call, the caller introduces the reason for the call in next turn. Sometimes the identity check is not focusing on the name of the callee, but on his tendency to be the right person to speak with in the context of the type of offer or proposal which is about to be made.

Fragment 7.

1. Lenny: hello: eh eh this is lenny!
2. Brianna: I need to speak to the person that handles the clp electric bill, is that you?
3. Lenny: hi uh uh ss- sorry, I'b- (0.3) I can barely hear you there?
4. Brianna: I need to speak to the person that handles the clp bill - electric bill - is that you?
5. Lenny: °ye-° ye:s yes yes.
6. Brianna: hello, my name is brianna, I'm calling on behalf of spark energy in regards to the connecticut electric choice program. do you recall receiving that information, sir?

The identification check, which has been introduced in her first turn by Brianna (Fragment 7), the caller, aims at finding the right person who is responsible for some task (here the electric bill). She repeats the same question after the hearing trouble question from Lenny. In this sequential context, the "Yes" turn displays a positive answer to the identification question. This understanding is embedded in how Brianna is pursuing the call with the reason for the call. No doubt that Lenny is the right addressee.

Fragment 8.

1. Lenny: hello: eh eh this is lenny!
2. Brian: hi lenny this is brian security specialist how are you today.
3. Lenny: hi uh uh ss- sorry, I'b- (0.3) I can barely hear you there?
4. Brian: my name is brian it's a pleasure to make your acquaintance lenny how'r yu
5. Lenny: °ye-° ye:s yes yes.
6. Brian: have you ever had a security system for yourself.
7. Lenny: oh good yes yes yes.
8. Brian: do you have a security system ?
9. Lenny: uh yes, yes, uh::uh, someone, someone did say last week or some- one did call last week about the same (.) thing, wa-was that, was that, you?
10. Brian: it wasn't me it might have been someone else to my company or something, .hh but do you need do you have a security system.

In such a sequential structure, the identity check or other verification questions ("Are you in front of your computer? Do you have a security system?") can be built as pre-sequences, which will sometimes freeze the introduction of the reason for the call. In Fragment 8, after the presentation and the "how are you" turn (T.2, 4), the caller introduces a verification question which is supposed to preface the offer. The telemarketer tries to ask Lenny whether he has a security system (T.6), but does not accept Lenny's the second enthusiastic "yes" turn (T.7) as a proper answer. Then, the telemarketer repeats the question (T.8). The next Lenny's turn, which begins with a "yes", could have been a second possible acceptable answer to the question, but the telemarketer keeps repeating the question (T.10). The several repeats of the same question display that there is an incoming issue in the conversation which has been noticed by the caller.

Nevertheless, such instances are very rare in the corpus. In most cases, Lenny does the job and the reason for the call can be introduced. The five first turns adjust to the various different sequential openings which have been found and get different senses from their positions in these sequential environments.

6. DISCUSSION

Lenny's efficiency is not only related to the specific design of Lenny's turns, but also to the orientations displayed by the caller in his proper turns. The caller's turns display his local understanding of Lenny's turns and he treats Lenny's turns as displaying some understanding of his own contributions. To a certain extent, it does not matter that Lenny's turns are fixed, pre-recorded items, as long as this feature is not discovered by the caller himself during the conversation. The practical sense of each turn at talk, either Lenny's ones or the caller's contributions, is embedded into the meaningful web of the call in progress.

A conversation analytic perspective on Lenny's calls reveals that the smartness of a bot can not be hidden in a sophisticated AI but in its tendency to participate to the sequential development of the relative diversity of calls without "freezing" the call. We have shown that this tendency is based both on the specific design of Lenny's turns and in their capacity to merge with the various sequential environments of different types of spam calls. We will complete in forthcoming papers this first study of openings with broader analysis of other sections of the same calls: the core parts of the calls, and the conversational treatment of the looping mode. Meanwhile, we would like to focus on the complexity of Lenny's character, which makes it difficult to replicate, while keeping its "botness" less visible for the caller.

6.1 Lenny the subtle bot

Lenny's talk displays a specific perspective which is very balanced in relation to the main orientations of the callers. Like other professional phone talk settings, unsolicited spam calls are script-guided and goal-oriented [27]. As Mazeland [52] has pointed out in one of the very few conversation analysis studies on telemarketing, the operators try to take control over the interaction with "initiatory actions" (i.e., first pair parts).

Accordingly, one of their first jobs is to check that Lenny can be correctly addressed as a member of a specific category (e.g., business owner) who is therefore entitled to [44,58,72] perform a specific activity (e.g., contracting a loan). Callers have little interest, if any, in addressing Lenny as an incumbent of other social categories (e.g., "grandfather") or collections of categories (such as "family"). For the same reason, callers are not "topically" oriented: they have no specific interest in "talking" about other topics that people usually used to bring into ordinary conversations.

Lenny's talk displays some features which foster callers: he is ready to talk; he displays some positive alignment in the very first turns to the reason for the call; he provides some confirmation of the requested identity. Then the callers have to deal with other aspects of Lenny's conduct, which complicates their job. First, they have to address the several repeat queries and verification questions from Lenny, without getting lost in the script that most operators hearably follow. So many repetitions tend to threaten the very work of turning the script that scammers use into the conversation. Repetition queries disturb the organization of the script: some callers used to jump to the next scripted turn instead of repeating their previous turn. Second, callers have to find ways to deal with Lenny's narratives, which are centered on family matters. Either they display alignment as possible recipients to such narratives, or they keep some distance with them and try to come back to their business talk as soon as they can. Both repeat queries, confirmation queries, and self narratives allow Lenny to control the turn management and/or sequential progressivity. Such attempts are difficult to handle, because most callers share the same orientation to a scripted interrogative series through which they keep control over the conversation.

Lenny's efficiency is deeply rooted in its propensity to maintain such a balanced orientation towards the call. Lenny leads the callers to adjust their own talk to the specificities of callee's productions, while maintaining a continuing, positive orientation to the business of the call. Its brilliant design lies in the subtle equilibrium it preserves between control and alignment.[5]

6.2 Usability of Transferring Calls to Lenny

In this paper we did not study the user aspect of transferring calls to Lenny. In fact, we have limited control and data on this aspect of the deployment, but in general the usability of the call transfer is quite poor. Requesting a user to perform multiple steps to transfer the call is not likely to scale well with the general public. On an enterprise desk phone where buttons can be configured to automatically transfer calls to a given phone number, the operation can be straightforward. On the other hand, such tasks are difficult to automate on mobile phones: call control APIs are very limited and the audio of a call is in general directly handled by the mobile baseband chip. As a consequence the audio stream is not easily accessible by applications on unmodified smartphones. Thus, automating the use of such chatbots with a smartphone application, without the involvement of an operator side telephony system, is currently very difficult to achieve. Nevertheless, the number of people using Lenny have been increasing as its popularity increases among the online community.

6.3 Comparing Lenny with Existing Voice Spam Countermeasures

Chatbots like Lenny does not necessarily prevent voice spam, in fact, using Lenny may increase the number of unwanted calls one receives, due to getting marked as a potential customer. In this respect, Lenny does not really compare with the other voice spam countermeasures that often aim to detect and block spam calls [69]. In fact, the recipient will still be disturbed with the call, and will need to make a decision on the call type (spam or not) to transfer the call. Moreover, the usability issues with call transfer and the possible need for a third party system reduces the scalability of such chatbots.

6.4 Effects on the Economics of Voice Spam

Lenny provides an opportunity to stall fraudsters and slow down economics of voice spam, by directly and indirectly increasing the cost of a failed telemarketing or scam call.

To spend 15 minutes or more of a working time with a Lenny-like bot represents a direct cost for spammers. More

[5] More Conversation Analysis work will be necessary to gain a proper understanding of the skilled Lenny.

importantly, it also results in an opportunity cost, because the spammer will not be able to target other legitimate customers during this time. This increases the call costs until reaching a valid customer and decreases the volume of calls a single spammer can generate in a certain time period [68]. On the other hand, victims could save time by using the chatbot instead of declining the proposal or dropping the call.

Depending on the expected monetary benefit of a spam scheme and the rate of use of chatbots, a spam campaign may become less profitable, or even not be economically viable. However, this would require a large number of chatbot users. In fact, a recent survey shows that more than 90% of participants do not listen to telemarketing proposals until the end; they either politely decline or hang up the call [24]. Another benefit of the generalization of such a service would be to reduce the economic damage of voice spam on society, both due to the direct monetary losses [48], and due to the reduced productivity [25].

A possible consequence is that the spammers will get acquainted with the chatbots and be able to quickly recognize and avoid them. Thus, a generic framework could be useful to simplify the creation of personal chatbots, e.g., providing guidelines on script preparation.

6.5 Guidelines for the design of Lenny-like bots

In the near future, we will try to develop the implications of our findings on the design of anti-scam chatbots thanks to a closer collaboration with their designers, either profane or professionals. For the time being, we propose some general guidelines for the design of such bots, based on our preliminary analysis of Lenny's usability:

- Maximize the coherence between all easily recognizable features of the chatbot which are available at first hearing: the voice, the local accent, the gender and the class of age membership have to be congruent in some way. Other category memberships can be revealed during the call: For instance, the callee can reveal that she is a "mother", a "daughter" or a "musician" during one of the narratives.

- The first available recognizable identity of the bot has to be tied, in one way or another, to the production of a series of specific type of turns: repeat queries. Design carefully a variety of repeat queries which can be based on different motives: hearing issues, connection problems, technical problems, incidents during the calls, interruptions from co-present others, etc.

- Design a list of queries checking the identity of the caller, the proper name of the institution he is calling on behalf, how much time he needs for this call, the precise nature of his firm's main activities, etc.

- Design three or four multi-unit turns. In each of these longer turns, the first unit which begins the turn has to display that the following turn will not be connected to the previous ones, using a "misplacement marker" (e.g., "by the way". See [60]). The following turn constructional unit will deliver a narrative about some event which is coherently tied to the first, recognizable, membership category of the caller or, on the contrary, which will add a new tied category membership. During the narrative, do not forget to design some short pauses after each main narrative component in order to invite the hearer to display some recipiency.

- Design an attention checking turn ("hello?" or "are you still there?") which will be activated after a few seconds of silence (the exact duration should be confirmed with a few tests) after any turn of the chatbot.

- Design carefully the sequential order and the design of the first turns, which will facilitate or block the initiation of the call and the introduction of the reason for the call.

- The script has to preserve an equilibrium between turns which project a next turn from the caller (first pair parts) and responsive turns which have to display a positive orientation to the previous, unknown caller turns.

- Record at least twenty turns, or more to prevent the risk of the looping mode, which may reveal that the callee is a bot.

This list of design proposals has been conceived from our efforts to understand the effectiveness of Lenny. Therefore, its purpose is to facilitate the design of Lenny-like bots to be used in the specific and limited context of scam calls, not to provide a series of rules for bot design. The efficiency of Lenny-like bots will rely on the unfolding course of each conversation and will rest on the situated understandings of the callers, who adjust their actions accordingly.

7. CONCLUSION

Voice spam is a prevalent, yet unsolved problem affecting telephone users. In this paper, we study a particular anti-spam chatbot, *Lenny*, which was created to fight such spam calls with a set of pre-recorded voice messages.

We first present several statistics showing that despite its simplicity, Lenny is very effective in dealing with phone spammers. Then, we propose to investigate the usability of Lenny from the perspective of Applied Conversation Analysis. We highlight the complexities of Lenny which are "seen but unnoticed" [34] by his co-conversationalists. Despite the apparent simplicity of this 16 pre-recorded turns chatbot, we show that its success relies on a sophisticated equilibrium between contrastive features: These features give it the necessary flexibility to fit into several sequential organizations, while keeping sufficient control over the interaction.

Our study also reveals various insights on the voice spam landscape and common strategies of phone spammers. Finally, we discuss several factors on the usability of chatbots against voice spam and possible effects on spam economics. We believe that widespread adoption of diverse chatbots can be effective in decreasing financial incentives of spam campaigns.

8. ACKNOWLEDGMENTS

This work was partially funded by the Principality of Monaco. We also thank the anonymous reviewers for their invaluable feedback, and the author of the telephony honeypot for sharing the call records with us.

9. REFERENCES

[1] Chatbot conference. https://chatbotconf.com/.

[2] How do I waste a telemarketers time? available at: http://no-more-calls.com/how-to-waste-a-telemarketers-time-2/.

[3] How does wasting a telemarketers time cost them money? available at: http://no-more-calls.com/how-to-waste-a-telemarketers-time/.

[4] Phone scam callers get owned! https://www.youtube.com/channel/UCxRY9vRnEfnijWJjfUE9xzQ.

[5] Taoa.net , Lenny! https://www.youtube.com/playlist?list=PLduL71_GKzHHk4hLga0nOGWrXlhl-i_3g.

[6] Suing a telemarketer: How I spent my summer vacation. available at: https://www.privacyrights.org, October 2007.

[7] Astycrapper v.03. available at: http://web.archive.org/web/20081030013832/http://www.linuxsystems.com.au:80/astycrapper/, October 2008.

[8] Lenny, the bot that tricks telemarketers, 2015. http://toao.net/595-lenny.

[9] Lenny's history & why he isn't creative commons, 2016. https://www.reddit.com/r/itslenny/comments/5lcfwq/.

[10] [NSFW] Telemarketer tells Lenny not to play with someone whose job it is to mess with him. available at: https://www.youtube.com/watch?v=GBSok8sPEM0, June 2016.

[11] Ray, Anna, Maya, and Tracy from Verde Energy spend an entire hour with Lenny. available at: https://www.youtube.com/watch?v=D-RAlSGWQlI&t=2s, September 2016.

[12] Top predictive dialers, cloud call centers, power dialers, auto dialers - Terminology. available at: https://www.telephonelists.biz/top-predictive-dialers-cloud-call-centers-power-dialers-auto-dialers/, 2016.

[13] Call center pricing. available at: https://www.worldwidecallcenters.com/call-center-pricing/, 2017.

[14] Jolly Roger Telephone Company. http://www.jollyrogertelco.com/, 2017.

[15] Lenny! available at: https://www.youtube.com/watch?v=Gj7AgYt4C6c&list=PLduL71_GKzHHk4hLga0nOGWrXlhl-i_3g, 2017.

[16] Outsource2india's call center pricing. available at: https://www.outsource2india.com/callcenter/pricing.asp, 2017.

[17] Phone fraud stops here. available at: https://www.pindrop.com/, 2017.

[18] Stop robocalls and telemarketers with Nomorobo. available at: https://www.nomorobo.com/, 2017.

[19] K. Armstrong. Conversational banking will transform the financial services industry. available at: https://thefinancialbrand.com/63772/conversational-banking-chatbots-bots-ai-messaging/, February 2017.

[20] V. A. Balasubramaniyan, A. Poonawalla, M. Ahamad, M. T. Hunter, and P. Traynor. Pindr0p: using single-ended audio features to determine call provenance. In *Proceedings of the 17th ACM conference on Computer and communications security*, pages 109–120. ACM, 2010.

[21] A. Banach. Outbound telemarketing strategies. available at: http://smallbusiness.chron.com/outbound-telemarketing-strategies-24269.html.

[22] D. Bolton. Meet Lenny - the Internet's favourite telemarketer-tricking robot. The Independent, Thursday 14 January 2016 http://www.independent.co.uk/life-style/gadgets-and-tech/news/lenny-telemarketer-bot-robot-prank-a6813081.html, 2016.

[23] R. Bostelaar. Lenny the call-bot tortures telemarketers: just ask the woman calling on behalf of Pierre Poilievre. available at: http://news.nationalpost.com/, August 2015.

[24] C. Brosset and G. Caret. Le demarchage telephonique et vous. available at: https://www.quechoisir.org, January 2017.

[25] E. Brown. Spam phone calls cost US small businesses half a billion dollars in lost productivity. available at: http://www.zdnet.com/article/spam-phone-calls-cost-us-small-businesses-half-a-billion-dollars-in-lost-productivity/, February 2014.

[26] A. Burlacu. AT&T starts robocalls crackdown with Call Protect service to block spam phone calls: How it works. available at: http://www.techtimes.com, December 2016.

[27] P. Drew and J. Heritage. In *Talk at work: interaction in institutional settings*. Cambridge University Press, 1992.

[28] Federal Communications Commission. FCC robocall and caller ID spoofing workshop, Sept 2015. Video recording available at https://www.fcc.gov/events/workshop-focus-robocall-blocking-and-caller-id-spoofing.

[29] Federal Trade Commission. Consumer Sentinel Network Reports. available at: https://www.ftc.gov/enforcement/consumer-sentinel-network/reports, 2008 to 2015.

[30] Federal Trade Commission. Robocalls: Humanity strikes back, 2015. Available at https://www.ftc.gov/news-events/contests/robocalls-humanity-strikes-back.

[31] Federal Trade Commission. National Do Not Call Registry Data Book FY 2016, December 2016.

[32] Federal Trade Commission. Q&A for Telemarketers and Sellers About DNC Provisions in TSR, August 2016.

[33] S. Gallagher. "You took so much time to joke me"-two hours trolling a Windows support scammer. available at: https://arstechnica.com, January 2017.

[34] H. Garfinkel. *Studies in ethnomethodology*. Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1967.

[35] P. Gupta, M. Ahamad, J. Curtis, V. Balasubramaniyan, and A. Bobotek. M3AAWG Telephony Honeypots: Benefits and Deployment Options, 2014.

[36] P. Gupta, B. Srinivasan, V. Balasubramaniyan, and M. Ahamad. Phoneypot: Data-driven understanding of telephony threats. In *22nd Annual Network and Distributed System Security Symposium, NDSS 2015, San Diego, California, USA, February 8-11, 2014*, 2015.

[37] J. Heritage. A change-of-state token and aspects of its sequential placement. In J. M. Atkinson, , and J. Heritage, editors, *Structures of social action: Studies in Conversation Analysis*, chapter 13, pages 299–345. Cambridge University Press, Cambridge, U.K., 1984.

[38] J. Heritage. Conversation analysis and institutional talk: analysing data. In D. Silverman, editor, *Qualitative research: Theory, method and practice*, pages 161–82. Sage Publications, London, 1997.

[39] J. Heritage. Oh-prefaced responses to inquiry. *Language in Society*, 27:291–334, 1998.

[40] J. Heritage. Oh-prefaced responses to assessments: A method of modifying agreement/disagreement. In C. E. Ford, B. A. Fox, and S. A. Thompson, editors, *The Language of Turn and Sequence*, pages 196–224. Oxford, 2002.

[41] J. Heritage. Commentary: On the diversity of 'changes of state' and their indices. *Journal of Pragmatics*, 104:207–210, 2016.

[42] S. Hester and P. Eglin. Membership categorization analysis: An introduction. In S. Hester and P. Eglin, editors, *Culture in action: studies in membership categorization analysis*, pages 1–24. University Press of America, Washington, D.C., 1997.

[43] W. S. Horton, D. H. Spieler, and E. Shriberg. A corpus analysis of patterns of age-related change in conversational speech. In *Psychology and Aging*, pages 708–713. 2010.

[44] W. Housley and R. Fitzgerald. The reconsidered model of membership categorization analysis. *Qualitative Research*, 2:59–83, 2002.

[45] Infinit Contact. Why is India losing 70% of call center business to the Philippines. available at: `http://www.infinitcontact.com/blog/india-losing-70-call-center-business-philippines/`, May 2014.

[46] G. Jefferson. Transcription notation. *Structures of Social Interaction*, 1984. A light version that we use is available at: `pages.ucsd.edu/~johnson/COGS102B/JeffersonianNotation.doc`.

[47] J. L. Josephson. The Economics of Business-to-Business (B2B) Telemarketing. document by JV/M Inc, available at: `http://www.jvminc.com/Clients/JVP/Economics.pdf`, 2011.

[48] K. F. Kok. Truecaller insights special report. available at: `https://blog.truecaller.com/2017/04/19/truecaller-us-spam-report-2017/`, April 2017.

[49] N. Levy. Amazon's $2.5M Alexa Prize seeks chatbot that can converse intelligently for 20 minutes. available at: `http://www.geekwire.com`, September 2016.

[50] A. Marzuoli. Call me: Gathering threat intelligence on telephony scams to detect fraud. Talk by Pindrop, available at: `www.blackhat.com`, August 2016.

[51] A. Marzuoli, H. A. Kingravi, D. Dewey, and R. Pienta. Uncovering the landscape of fraud and spam in the telephony channel. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 853–858, December 2016.

[52] H. Mazeland. Responding to the double implication of telemarketers' opinion queries. *Discourse Studies*, 6(1):95–115, 2004.

[53] H. Mazeland. Conversation analysis. *Encyclopedia of language and linguistics*, 3:153–162, 2006.

[54] N. Miramirkhani, O. Starov, and N. Nikiforakis. Dial one for scam: Analyzing and detecting technical support scams. *CoRR*, abs/1607.06891, 2016.

[55] N. R. Norrick. Interjections as pragmatic markers. *Journal of Pragmatics*, 41(5):866 – 891, 2009.

[56] G. Raymond. Grammar and social organisation: Yes/no interrogatives and the structure of responding. *American Sociological Review*, 2003.

[57] H. Sacks. An initial investigation of the usability of conversational data for doing sociology. In D. Sudnow, editor, *Studies in social interaction*, chapter 2, pages 31–74. Free Press, New York, 1972.

[58] H. Sacks. *Lectures on conversation*, volume 1-2. Basil Blackwell, Oxford, 1992.

[59] H. Sacks, E. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4, Part 1):696–735, December 1974.

[60] H. Sacks and E. A. Schegloff. Opening up closings. *Semiotica*, 8(4):289–327, 1973.

[61] M. Sahin, A. Francillon, P. Gupta, and M. Ahamad. Sok: Fraud in telephony networks. In *Proceedings of the 2nd IEEE European Symposium on Security and Privacy (EuroS&P'17)*, EuroS&P'17. IEEE, April 2017.

[62] E. A. Schegloff. Sequencing in conversational openings. *American Anthropologist*, 70(6):1075–1095, 1968.

[63] E. A. Schegloff. The routine as achievement. *Human Studies*, 9(2):111–151, 1986.

[64] E. A. Schegloff. On integrity in inquiry... of the investigated, not the investigator. *Discourse Studies*, 7(4-5):455–480, oct 2005.

[65] E. A. Schegloff, G. Jefferson, and H. Sacks. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382, 1977.

[66] A. Sobczak. 42 telesales, telemarketing, inside sales, and cold calling tips you can use right now to get more business and avoid rejection. available at: `http://businessbyphone.com/telemarketing-tips/`, 2015.

[67] J. L. Tabron. Linguistic features of phone scams: A qualitative survey. *11th Annual Symposium on Information Assurance (ASIA'16)*, June 2016.

[68] Talks To Telemarketers . Predictive Dialers and Robocalls are poor Marketing. document by JV/M Inc, available at: `http://www.tormentingtelemarketers.com/2015/09/predictive-dialers-and-`

robocalls-are-poor-marketing/, September 2015.

[69] H. Tu, A. Doupé, Z. Zhao, and G.-J. Ahn. SoK: Everyone Hates Robocalls: A Survey of Techniques against Telephone Spam. In *Proceedings of the 37th IEEE Symposium on Security and Privacy*, May 2016.

[70] J. Valentine. 3 reasons the call center is far from dead. available at: http://mashable.com/2012/04/24/call-center-death-exaggerated/#5o3Fc5GKiZqF, April 2012.

[71] Y. Wang. Your next new best friend might be a robot. available at: http://nautil.us/issue/33/attraction/your-next-new-best-friend-might-be-a-robot, February 2016.

[72] R. Watson. *Some General Reflections on Categorization and Sequence in the Analysis of Conversation*. Unversity Press of America, 1997.

[73] J. Weizenbaum. Eliza;a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45, January 1966.

APPENDIX

A. ROUGH TRANSCRIPT OF A TELEMARKETING CALL

[00:00:00] Lenny: hello: thi- this is Lenny!
[00:00:03] Telemarketer: lenny, i'm looking for mr. [00:00:04 sound cut]
[00:00:06] Lenny: uh-- sso- sorry, I'b- I can barely hear you there?
[00:00:13] Telemarketer: homeowner.
[00:00:15] Lenny: ye- yes yes yes.
[00:00:19] Telemarketer: mr. [00:00:19 sound cut] we're giving free estimates for any work you need on your house. were you thinking about having any projects? a little craning driveway, roof work, anything you need done. we'll give you a free estimate.
[00:00:31] Lenny: oh good, yes, yes, yes.
[00:00:34] Telemarketer: what would you like to have done? what were you thinking about? anything around the house?
[00:00:39] Lenny: uh yes, yes, uh::uh, someone, someone did- did say last week or some- one did call last week about the same (.) thing, wa-was that, was that, you?
[00:00:50] Telemarketer: no, sir. i've might have been in another company. what was it that you were doing?
[00:00:55] Lenny: ye-yes. ss- sorry, what- wa- what was your name again?
[00:01:00] Telemarketer: yes. what were you thinking about having done?
[00:01:04] Lenny: well, it- it it's funny that you should call, because, my third eldest larissa, uhh, she, she was talking about this. (.) u:h #just this last week and .hh you you know, sh- she is-, she is very smart, I would- I would give her that, because, you know she was the first in the family, to go to the university, and she passed with distinctions, you know we're- we're all quite proud of her yes yes, so uhh:: yes she was saying that I should, look, you know, get into the, look into this sort of thing. uhh so, what more can you tell me about it ?
[00:01:14] Telemarketer: #mm-hmm. okay. alright. well, good, good. [inaudible 00:01:33] so you're very proud. okay.# well, we are full-service construction company. we do everything from the roof to the foundation. we've been in business for over 32 years. we're licensed, bonded, and insured, and we have plenty plenty of references if you need them. where you thinking about doing any work inside or outside?
[00:02:05] Lenny: I: I am sorry, I, I (.) couldn't quite catch you th-, catch you there. wha- what was that again?
[00:02:12] Telemarketer: where you thinking about doing work inside or outside?
[00:02:17] Lenny: uh. ( ) the: (.) ss sorry, aw again?
[00:02:23] Telemarketer: [laughs] where you going to do work inside or outside?
[00:02:28] Lenny: cou-could you say that again- again please?
[00:02:32] Telemarketer: i tell you what, i'm going to send one of my guys over to your place. you're at six [00:02:37 sound cut]. he can sit down with you. he'll discuss everything about our services. he'll give you our coupon. it's up to 50% off. i'll have him there, let's see it's 12:30, i can get him over there by 2:30. are you and your wife be home at 2:30? we'll come by, show show you all our stuff and you can let us know what you wanna do then. okay.
[00:02:59] Lenny: yes, yes, yes...
[00:03:01] Telemarketer: that makes sense?
[00:03:03] Lenny: sorry uhh, which company did you say you were calling from again?
[00:03:08] Telemarketer: wise. w-i-s-e. it'd be very wise for you to use our services. that's our commercial.
[00:03:15] Lenny: well, you know. here's- here's the thing because the last time that I--that someone called up, uh #and spoke to me on the phone, I got in quite a bit of trouble from--with the people here because I went for something that I shouldn't have. uh, I probably shouldn't be-be telling you that. but um, yes, I-I think m-my- my eldest Rachel, she-she uh, uh would-wouldn't speak to me for a week, now, you know that-- that happens, you know but uh it bit--that really hurt and-and-and sometimes in the family you know these-these things are quite important you know. they're more important than uh--uh any, you know, job or-or-- Phone call or-or-- what- wha-- whatever it is.
[00:03:22] Telemarketer: #mm-hmm. umm-hmm. umm-hmm. that's okay? mm-hmm. oh boy she got mad at you?# of course, family is always important. now let me asked you. uh, is three o'clock going to be good for you and your wife?
[00:04:11] Lenny: well yeah, since--since you-you put it that way, I mean you-you've been quite friendly and straightforward with me here. #um, h-hello?
[00:04:20] Telemarketer: #great. very good.# yes, i'm here. thank you.
[00:04:25] Lenny: hello? #are you there?

[00:04:25] Telemarketer: #i'm just saying, thank you.# yes sir, i'm here.
[00:04:30] Lenny: #oh yes-- s-sorry. Is-is--I have a-- have a bit of a--bit of a problem with this phone-- --and-and my hearing is not so good. um, yes,# uh, w-wha- sorry, wha-what were you saying again?
[00:04:30] Telemarketer: #hello? that's okay. that's okay. no problem.# i just saying that it was a pleasure speaking with you as well and we're going to have my guy come out and talk to you and your wife about three o'clock. i just wanna be able to let him know what it is that you guys were thinking about doing on the house. was it painting or kitchen, bathroom remodeling? what was it that you guys wanted to have look at?
[00:05:05] Lenny: well, you know with-with the world finances the way they are I know you know we're not-we're not allowed to spend as much as-as what we were. [stammering] how-how:: how-how-how is this going to uhh h-how is this going to work?
[00:05:23] Telemarketer: well, we'll have to come out and see what the job is first before we could talk about any form of uh, money but, uh, you won't have to worry about that until we see what it is that you need done.
[00:05:36] Lenny: #well,that-that-that does sound good. I mean, you-you have been very patient with an old man here and uh [laugh] Bt-it's uh-- yeah I mean, uh, it's-it's something that-that I've been told that I should be looking at-- uh-, my third eldest laris-larissa, she uh, I think I mentioned larissa before (.) yes-yes she uh-- she says th-that I should be going for the--something like this but uh, it's just a matter of what , you know, what-what is most appropriate for--for uhh th-the time and I guess what not. sorry could you-- just hang on# for one second here? hang on. [ducks quacking in the background]
[00:05:36] Telemarketer: #what it is that you are thinking of doing? oh, no problem , that's -- that's our job. i mean your purpose so. yes what was she said-- she told you to have done? mm-hmm-- i got you.#
[00:06:32] Lenny: yeah. so-sorry #about that. uh-- s-sorry# wha-what were you saying there again?
[00:06:33] Telemarketer: #yes, sir. that's okay.# i was asking what work that you need done?
[00:06:43] Lenny: uh yes, yes, uh::uh, someone, someone did- did say last week or some- one did call last week about the same (.) thing, wa-was that, was that, you?
[00:06:54] Telemarketer: no sir. that may have been another company.
[00:06:58] Lenny: ye-yes. ss- sorry, what- wa- what was your name again?
[00:07:03] Telemarketer: my name is michael.
[00:07:06] Lenny: well, it- it it's funny that you should call, because, my third eldest larissa, #uhh, she, she was talking about this. (.) u:h just this last week and .hh you you know, sh- she is-, she is very smart, I would- I would give her that, because, you know she was the first in the family, to go to the university, and she passed with distinctions, you know we're- we're all quite proud of her yes yes, so uhh:: yes she was saying that I should, look, you know, get into the, look into this sort of thing. uhh so, what more# can you tell me about it ?
[00:07:12] Telemarketer: #mm-hmm, mm-hmm-- what was she talking about? right. what was she talking about? what was she talking about? what was she talking about? what were she talking about mr. [00:07:32 sound cut]? looking to what sort of thing mr. [00:07:43 sound cut] ?# what would she like to have done mr.[00:07:49 sound cut]?
[00:07:51] Lenny: I: I am sorry, I, I (.) couldn't quite catch you th-, catch you there. wha-what was that again?
[00:07:57] Telemarketer: what do you want done?
[00:08:00] Lenny: uh. ( ) the: (.) ss sorry, aw again?
[00:08:06] Telemarketer: well, i guess we'll gonna be here a while. what did she want done?
[00:08:11] Lenny: cou-could you #say that again- again please?
[00:08:12] Telemarketer: #i mean bathrooms.# so do you need your bathroom redone?
[00:08:19] Lenny: #yes, yes, yes...#
[00:08:19] Telemarketer: #maybe your kitchen# how about the drive way? maybe even the garage? have you done any work on your roof?
[00:08:27] Lenny : sorry uhh, which company did you say you were calling from again?
[00:08:32] Telemarketer: i didn't say, uh, the thing is we were tying to see what did you need done.
[00:08:39] Lenny: #well, you know. here's- here's the thing because the last time that I-- that someone called up, uh and spoke to me on the phone, I got in quite a bit of trouble from--with the people here because I went for something that I shouldn't have. uh, I probably shouldn't be-be telling you that. but um, yes, I-I think m-my- my eldest Rachel, she-she uh, uh would-wouldn't speak to me for a week, now, you know that-- that happens,

you know but uh it bit--that really hurt and-and-and sometimes in the family you know these-these things are quite important you know. they're more important than uh--uh any, you know, job or-or-- Phone call or-or-- what- wha-- whatever it is.#
[00:08:39] Telemarketer: #although i love having this conversation. i get paid by the hours, so the longer i sit, the longer i talk with you, the better um, yeah, right. um, um, how often do you do this? [laughs] this is so much fun. i-- i've never seen anybody have their own routine over the phone. this is quite cool since both of us are going to talk. now i'm thinking this is maybe recording because you can't hear anything that i'm saying to you at this point. so we might as well just go ahead and do this over.# so now you're gonna ask me, "what did i say? i didn't hear you. would you repeat that?"
[00:09:31] Lenny: well yeah, since--since you-you put it that way, I mean you-you've been quite friendly and straightforward with me here. um, h-hello?
[00:09:45] Lenny: hello? are you there? oh yes-- s-sorry. Is-is--I have a-- have a bit of a-- bit of a problem with this phone-- --and-and my hearing is not so good. um--#um, yes, uh, w-wha- sorry, wha-what were you saying again?
[00:09:59] Telemarketer: #i ran into a building, that's not--# did you hear them?
[00:10:15] Lenny : well, you know with-with the world finances the way they are I know you know we're not-we're not allowed to spend as much as-as what #we were. [stammering] how-how:: how-how-how is this going to uhh h-how is this going to work?
[00:10:29] Telemarketer: #[laughs] this is great.#
[00:10:30] Lenny: h-how is this going to work? hello? are you there?uh yes-- s-sorry wha-what were you saying there again?
[00:11:16] [END OF AUDIO]

B. ROUGH TRANSCRIPT OF A SCAM CALL

[00:00:00] Lenny: hello: thi- this is Lenny!
[00:00:04] Adam: yeah mr. lenny, you have been chosen to get a lower interest rate, so i believe you have pressed one to get a lower interest rate right?
[00:00:13] Lenny: uh-- sso- sorry, I'b- I can barely hear you there?
[00:00:17] Adam: i'm saying so i believe you have pressed one to get a lower interest rate, right?
[00:00:24] Lenny: ye- yes yes yes
[00:00:26] Adam: okay, the interest you're paying at the moment is 19.9, right?
[00:00:32] Lenny: oh good, yes, yes, yes.
[00:00:34] Adam: and we are going to drop that down to less than 10% on this same call okay?
[00:00:40] Lenny: uh yes, yes, uh::uh, someone, someone did- did say last week or some- one did call last week about the same (.) thing, wa-was that, was that, you?
[00:00:50] Adam: oh okay, and did they provide you the low interest?
[00:00:56] Lenny: ye-yes. ss- sorry, what- wa- what was your name again?
[00:01:01] Adam: sir i'm saying my name is adam, adam chaw and i'm saying did they provide you the lower interest?
[00:01:09] Lenny: well, it- it it's funny that you should call, because, my third eldest larissa, uhh, she, she was talking about this. (.) u:h just this last week and .hh you you know, sh- she is-, she is very smart, I would- I would give her that, because, you know she was the first in the family, to go to the university, #and she passed with distinctions, you know we're- we're all quite proud of her yes yes, so uhh:: yes she was saying that I should, look, you know, get into the, look into this sort of thing. uhh so, what more can you tell me about it ?
[00:01:29] Adam: #yeah# so as you know today you are getting this call from low interest rate department working for the head office of visa and mastercard and you have been chosen only because of your good payment history. for the past six to seven months, you have been making your payments on time, right? you always try to make more the minimum payments right?
[00:02:10] Lenny: I: I am sorry, I, I (.) couldn't quite catch you th-, catch you# there. wha -what was that again?
[00:02:13] Adam: #you always try to make more than # the minimum payments, right?
[00:02:18] Lenny: uh. ( ) the: (.) ss sorry, aw again?
[00:02:23] Adam: you always try to make more than the minimum payments, correct sir?
[00:02:28] Lenny: cou-could you say that again- again please?
[00:02:31] Adam: sir, i'm asking you, you always try to make your payments on time, right?
[00:02:37] Lenny: yes, yes, yesâĂę
[00:02:39] Adam: okay, and today that's the reason you're getting this call and that's the reason we are going to provide to lo--lower interest rate because of your good payment

history, okay.

[00:02:50] Lenny: sorry uh, which company did you say you were calling from again?

[00:02:54] Adam: sir, we are working for the head office of visa and mastercard, working with the head office of visa and mastercard and that's the reason we are going to provide you the low interest, okay. so grab your card on hand and verify me the membership number starting from five.

[00:03:09] Lenny: well, you know. here's- here's the thing because the last time that I--that someone called up, uh and spoke to me on the phone, I got in quite a bit of trouble from --with the people here because I went for something that I shouldn't have. uh, I probably shouldn't be-be telling you that. but um, yes, I-I think m-my- my eldest Rachel, she-she uh, uh would-wouldn't speak to me for a week, now, you know that-- that happens, you know# but uh it bit--that really hurt and-and-and sometimes in the family you know these-these things are quite important you know. they're more important #than uh--uh any, you know, job or-or-- Phone call or-or-- what- wha-- whatever it is.

[00:03:40] Adam: #you tell me your eldest--the daughter's name for some correction, yeah mr. lenny i understand that, i understand mr. lenny, that today we are going to provide you the lower interest # on this same call, so i need you to grab your mastercard on hand and verify me the membership number starting from five, can you do that?

[00:04:07] Lenny: well yeah, since-#-since you-you put it that way, I mean you-you've been quite friendly and straightforward with me here. um, #h-hello?

[00:04:08] Adam: #can you grab you card and verify me the membership number? # yes, yeah sir.

[00:04:20] Lenny: hello? are you there?

[00:04:24] Adam: yes sir, i'm here. grab #your card and verify me the membership number starting from five.

[00:04:26] Lenny: #oh yes-- s-sorry. Is-is--I have a-- have a bit of a--bit of a problem with this phone-- --and-and my hearing is not so good. #um, yes, uh, w-wha- sorry, wha-what were you saying again?

[00:04:34] Adam: #[laugh] no problem, no problem.# grab your card sir, your mastercard and verify me the membership number starting from five.

[00:04:47] Lenny: well, you know with-with the world finances the way they are I know you know we're not-we're not allowed to spend as much as-as what we were. #[stammering] how-how:: how-how-how is this going to uhh h-how is this going to work?

[00:04:56] Adam: #yeah sir, i understand, i understand that completely and that's the reason i want to provide you the lower interest on your mastercard. #sir can you grab your mastercard?

[00:05:07] Lenny: well, that-that-that does sound good. I mean, you-you have been very patient with an old man here and uh [laugh] it-it's uh-- yeah I mean, uh, it's-it's something that-that I've been told that I should be looking at-- #uh-, my third eldest laris-larissa, #she uh, I think I mentioned larissa before (.) yes-yes she uh-- she says th-that I should be going for the--something like this but uh, it's just a matter of what , you know, what-what is most appropriate for--for uhh th-the time and I guess what not. sorry could you-- just hang on for one second here? hang on. [ducks quacking in the background]

[00:05:19] Adam: #okay, okay yeah so are you grabbing you card sir or should i hang up?#

[00:05:34] [END OF AUDIO]

Valuating Friends' Privacy: Does Anonymity of Sharing Personal Data Matter?

Yu Pu
Security, Privacy and
Information Economics Lab (SPIEL)
The Pennsylvania State University

Jens Grossklags
Chair for Cyber Trust
Department of Informatics
Technical University of Munich

ABSTRACT

Through their third-party app installation decisions, users are frequently triggering interdependent privacy consequences by sharing personal information of their friends who are unable to control these information flows. With our study, we aim to quantify the value which app users attribute to their friends' information (i.e., value of interdependent privacy) and to understand how this valuation is affected by two factors: sharing anonymity (i.e., whether disclosure of friends' information is anonymous), and context relevance (i.e., whether friends' information is necessary for apps' functionality). Specifically, we conduct a between-subject, choice-based conjoint analysis study with 4 treatment conditions (2 sharing anonymity $\times$ 2 context relevance). Our study confirms the important roles that sharing anonymity and context relevance play in the process of interdependent privacy valuation. In addition, we also investigate how other factors, e.g., individuals' personal attributes and experiences, affect interdependent privacy valuations by applying structural equation modeling analysis. Our research findings yield design implications as well as contribute to policy discussions to better account for the problem of interdependent privacy.

1. INTRODUCTION

The vast majority of published research on privacy-decision making focuses on individual choices regarding personal privacy. However, with the accelerating usage of Social Network Sites (SNSs), mobile platforms and other digital advances with interactive tools, we observe the increasing relevance of decisions which affect others' information. These *interdependent privacy* choices involve scenarios in which a decision-maker has power over the sharing of personal information about other individuals, which are often friends, family members or colleagues. Previous work has studied this problem space from a theoretical [11, 80] and behavioral perspective [83]. A key finding is that individuals exhibit behaviors which can be interpreted as *privacy egoism*: they value their own information much higher than the information of a friend [83]. From a theoretical perspective, this phenomenon can be explained with the economic concept of negative externalities, i.e., individuals do not bear the (privacy) cost that they impose on others [11].

However, the understanding of important contextual factors that influence interdependent privacy decision-making is still in its infancy. In particular, we do not yet understand how characteristics of the platform, which mediates the sharing, influence human choices about others' privacy. A key aspect is to which degree transparency (between the sharer and the affected individuals) about a sharing decision influences the propensity to share information, or affects valuation of personal information of friends. In other words, our central research question is whether different modes of *anonymity* (or identifiability) influence how a sharing decision is perceived, when it affects interdependent privacy valuation.

We consider the scenario of third-party app adoption on SNSs where users are presented with app offers and associated authorization dialogues which may trigger sharing decisions over their own personal information and their friends' personal information [107]. For example, an app may request to access not only users' own data, but also information about their friends. In practical settings, the ability of an affected individual to learn about others' sharing decisions is quite modest. For example, users may be subjected to social app advertisements and may indirectly learn that a friend has adopted an app which triggers the sharing of friends' information.[1] We focus on studying the impact of this veil of anonymity (as well as its counterpart full identifiability) of sharing decisions.

To address our research question, our first step is to quantify the interdependent privacy value by applying the methodology of conjoint analysis. In our previous work [83], we conducted a full-profile conjoint study to determine this value, and we use this study setup as a starting point for the current investigation. However, due to a high cognitive challenge presented by the full-profile method, alternative approaches should be taken to address low data quality (see Appendix A). To respond to this data quality concern and to improve on our previous work, we utilize a different methodology, i.e., choice-based conjoint analysis, to determine interdependent privacy valuations. Further, we introduce, in the choice-based conjoint study, four treatment scenarios which differ in whether or not sharing friends' data is anonymous (*sharing anonymity*[2]) and whether or not the requested friends' data is useful to app's functionalities (*context relevance*). This allows us to examine how sharing anonymity and context relevance affect app users' valuation towards their friends' data.

[1] In the mobile app context even such spurious cues may not exist when a user shares an address book or other data type containing friends' data. Likewise, in the context of genetic privacy there is no mechanism that automatically informs other family members about the decision by one individual to take a test [109].

[2] To clarify, sharing anonymity does not mean that an app user shares anonymized friends' data. Instead, it indicates the situation where it is hard for friends to identify the person that released their information to apps.

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

In order to comprehensively explain the valuation of friends' information, our second step is then to apply Structural Equation Modeling (SEM) analysis to investigate how interdependent privacy values are influenced by factors such as *other-regarding preferences* (see details in later sections), privacy knowledge, privacy concern, and the treatment conditions, i.e., sharing anonymity and context relevance.

Our results suggest that valuation of interdependent privacy is affected not only by individuals' personal attributes and experiences, such as other-regarding preferences and privacy knowledge, but also by treatment conditions. In particular, we find that anonymity plays an important role in interdependent privacy valuation. Specifically, when individuals believe the sharing of friends' information is anonymous, they tend to value their friends' data significantly *less*. Similarly, we find app users place a significantly lower value on their friends' information when they believe such information is useful for an app's functionality.

These results offer valuable insights into the problem of interdependent privacy, which are directly applicable to privacy by design or re-design initiatives [41, 116]. More specifically, our study conveys that design features helping to raise individuals' interdependent privacy concerns will also impact individuals' valuation of friends' personal information. But additional ways to protect friends' privacy emerge which can be used by interface designers and information architects. First, making the sharing of friends' data identifiable is a viable approach to erect a psychological hurdle against unfettered bulk data sharing with third parties as often triggered by app adoption. Second, informing app users when data collection is not contextually relevant also influences privacy valuations significantly. Computer scientists work on automating the analysis of contextual relevance in the app context by identifying over-privileged apps [28, 32, 46], which makes the implementation of related design features during the app selection process viable. In addition, our research findings also emphasize the important roles of governmental interventions and privacy education in protecting friends' privacy in the context of app adoption.

Roadmap: We proceed as follows. In Section 2, we discuss related work on the role of anonymity in individual decision-making. We further summarize extant work on the value of personal information, and the modeling of privacy decision-making. In addition, we also review existing work on resolving interdependent privacy conflicts. Next, we present the choice-based conjoint analysis approach, and the associated results in Section 3. In Section 4, we discuss the development and results for the behavioral model based on SEM. Finally, we discuss our findings in Section 5, and offer concluding remarks in Section 6.

2. RELATED WORK

2.1 Anonymity in Individual Decision-Making

A set of studies in the area of experimental economic research has focused on the influence of anonymity on decision-making. In particular, the experimental literature on economic bargaining games which mostly centers on the analyses of the so-called ultimatum [42] and dictator games [60] is of high relevance. In the classical version of both games, a monetary amount (i.e., pie) is offered for allocation between two individuals. One person acts as the proposer and can suggest a split of the pie. In the ultimatum game, the recipient of the proposal can reject the offer (then the money will remain with the experimenter) or accept the split [42]. In contrast, in the dictator game the recipient has no decision-making power (and the pie is allocated according to the proposed split) [60]. A specific sub-area of this literature is addressing the impact of anonymity from two perspectives: 1) anonymity between proposer and recipient, 2) anonymity between players and experimenter (i.e., double-blind).

Radner and Schotter compare face-to-face (F2F) bargaining with anonymous bargaining and find that the latter was associated with an increase in rejected proposals, while the former was associated with an almost uniform acceptance rate [85]. Prasnikar and Roth report similar results [78]. However, they also find that F2F communications that explicitly exclude any form of conversation about the relevant bargaining aspects and are merely social in nature, also contribute to an almost uniform acceptance rate of proposals which were later issued without additional F2F exchanges [78]. During the latter treatment, participants were required to learn the name and education level of their bargaining opponents. The finding of this social conversation treatment was interpreted to confirm that social pressures arising from F2F are influencing subjects; rather than the discussion of any pertinent aspects of the transaction [89]. Similarly, Charness and Gneezy conduct dictator and ultimatum game experiments in which they compare treatments in which participants were informed about the family names of their counterparts (or not) [18]. This manipulation strongly impacted the generosity of proposers in the dictator game, but not the initial offer of the proposers in the ultimatum game where strategic considerations seemed to prevail [18]. Hoffman et al. introduced a double-blind setup in which the experimenter could not identify the experimental participants [48]. The results indicate that this double-blind setup was associated with the most selfish offers by the proposers. Experiments have also been conducted in the field to document the negative impact of anonymity on donations for environmental causes [5] or in churches [96].

In addition, a stream of research in the field of information system investigates the impact of anonymity on individuals' self-disclosure on social network sites. These studies mainly focus on two types of anonymity: discursive anonymity and visual anonymity. Discursive anonymity refers to the extent to which information can be linked to a particular source [92], whereas visual anonymity indicates the degree to which others can see and/or hear the person who discloses the information [92]. Although focusing on this topic for more than a decade, researchers have not reached an agreement on either the impact of discursive anonymity or the influence of visual anonymity on self-disclosure. For example, Qian and Scott [84] report a positive relationship between self-disclosure and discursive anonymity. However, this association is found to be negative by Hollenbaugh and Everett [49]. When it comes to visual anonymity, some studies claim that it is positively related to self-disclosure [59, 49], while other research fails to detect such an association [84]. These contradictory empirical findings suggest that the relationship between anonymity and self-disclosure in online social networks is still in question and should be further examined [49].

Closely related to our work, some studies explore the impact of anonymity on individuals' privacy attitudes or privacy behaviors. In particular, through an empirical study, Jiang et al. [57] report that when individuals perceive themselves to be unidentifiable, they feel less concerned about their privacy. In addition, they find that individuals exhibit higher levels of concern about their own privacy when other parties' identities are anonymized. However, we are still unaware of any research that directly addresses how anonymity impacts individuals' attitudes towards others' privacy. Our study addresses this literature gap by exploring the impact of anonymity on the valuation of interdependent privacy.

2.2 Economic Value of Privacy

Several research projects explore the value of personal information which is also a central aspect of our study. For example, situating individuals in a second-price auction scenario, Huberman et al. elicit monetary valuations for individuals' weight and height (which were previously recorded) [54]. They find that deviation from a perceived population standard drives higher valuations. A similar approach is used to derive how individuals value information about their location traces [24].

In various experiments, participants are presented with situations in which they trade off privacy for better recommendations [97], a discount on a product purchase [3, 105], or a pure monetary reward [79]. In contrast, the willingness to pay to protect information, or to purchase a higher-priced option with privacy-friendlier terms is often reported to be low [9, 38, 79], though there are exceptions [31]. Other research shows that individuals associate little value with their own information on social network sites (when faced with the threat of deletion) [7].

Taking a different viewpoint, Acquisti and Grossklags as well as Grossklags and Barradale investigate how privacy and security preferences relate to financial preferences (e.g., discounting behaviors) [1, 39]. Böhme and Grossklags study how privacy norms shift on a micro-lending platform, when platform mechanisms for borrower-lender matching shift away from placing focus on personal descriptions provided by borrowers [12].

Several researchers have conducted studies with conjoint analyses beginning with Hann et al. [44, 43]. More recently, Krasnova et al. used this methodology to understand privacy concerns in SNS settings [62]. Common to these works is that they also determine the monetary value of personal information.

In our previous work, we determined the monetary value of personal and interdependent privacy by applying the conjoint study methodology [81, 83, 82]. We replicate the utilized basic scenario to conduct the conjoint study in this paper [83], however, we add as an additional treatment condition whether (or not) anonymity of sharing decisions over friends' information is provided. In addition, we utilize a different methodology by applying choice-based conjoint analysis to address data quality concerns from the full-profile method applied in our previous work [83].

2.3 Explaining Value of Privacy

In our work, we also aim to develop an explanatory model for the valuation of personal and interdependent privacy as measured with the conjoint study methodology.

Several related studies have focused on utilizing concerns for personal privacy as a key construct which is also part of our explanatory model [94]. We also draw on published works about antecedents of personal privacy concerns, for example, research which utilizes past privacy invasions as an explanatory factor for personal privacy concerns [95].

A number of models focus on effects of personal privacy concerns, for example, by trying to explain purchase intentions [30] or disclosure behaviors [65]. In contrast, our paper is focused on explaining personal and interdependent privacy *valuations* as in our previous work [83]. However, our behavioral model substantially differs by considering the explanatory effects of other-regarding preferences [22], perceived control [27], and disposition to value privacy [115]; in addition, our center of interest is to build a model to explain different treatment conditions regarding anonymity. We will detail the building blocks of our model in Section 4.1.

2.4 Resolve Interdependent Privacy Conflicts

Privacy conflicts may arise in interdependent privacy scenarios, where privacy preferences of those who share others' data and those whose information is leaked are not aligned. These privacy conflicts are referred to as multi-party privacy conflicts (MPCs) [103]. Several research projects explore how to resolve conflicts arising from interdependent privacy issues in social media, although not in the scenario of social app adoption. A stream of these studies focuses on providing computational mechanisms or external tools to deal with MPCs. For example, in the scenario of photo sharing on social network sites, a system has been proposed so that when a user is tagged in a photo, he/she can send privacy suggestions or feedback to those who upload the photo [10]. Also in the scenario of photo sharing, Ilia et al. [56] introduce a mechanism of blurring faces of individuals (who appear in photos) based on a users' access control permissions.

To provide support for users to resolve MPCs, some studies propose sharing policies based on aggregated individual privacy preferences. For example, Hu et al. [51] formulate an access control model, multi-party policy specification scheme, and a policy enforcement mechanism to facilitate collaborative management of shared data. Thomas et al. [103] demonstrate how Facebook's privacy model can be adapted to enforce multi-party privacy. Similarly, other mechanisms or access control policies have been introduced in [15, 101] to address MPCs.

Other researchers try to study MPCs from the perspective of game-theoretic analysis. For example, Hu et al. [52] study a multi-party access control model to investigate systematic approaches to identify and to resolve conflicts of collaborative data sharing. Similarly, a negotiation mechanism is introduced and examined to help users to reach an agreement in scenarios with MPCs [102].

There is another stream of studies which explores strategies users have utilized to resolve MPCs. Wisniewski et al. [112] demonstrate that individuals use both online strategies, such as untagging, and offline strategies, such as negotiating offline with affected others, before posting photos. In addition, they also investigate how support mechanisms that are provided by social media interfaces are used by individuals for addressing MPCs [111]. They conclude that these mechanisms are ineffective, difficult to use, and not easy to be aware of, and therefore users are more likely to apply offline coping strategies.

Conducting a qualitative study with 17 individuals, Lampinen et al. [64] discover that users apply a range of preventive strategies to avoid causing problematic situations for others. In particular, they categorize 4 types of strategies: preventive, corrective, individual, and collaborative. Similarly, Cho and Filippova [20] identify the same types of strategies based on findings from focus-group interviews and online surveys.

In practice, we would expect that individuals would regret many app adoption decisions, when they revisit apps' privacy practices or suffer from conflicts with their friends [34].

Finally, outside the context of SNSs, Harkous and Aberer investigate sharing practices on cloud storage platforms involving the access of third-party cloud storage apps to users' data repositories by conducting measurements and user studies [45].

In aggregate, most of these studies investigate different ways of resolving privacy conflicts that arise from interdependent privacy issues in social networks. However, we are unaware of any research that directly explores MPCs in the context of social app adoption. Our research provides insights for dealing with such privacy conflicts.

3. CONJOINT ANALYSIS TO DETERMINE PRIVACY VALUE

3.1 Design of Choice-based Conjoint Study

Conjoint analysis is a general approach for analyzing consumer preferences for multi-attribute products and services [37]. In a conjoint analysis study, it is often assumed that consumers view a product as a bundle of certain features (*attributes*), which have different values (*levels*) [36]. Through testing and analyzing individuals' preferences for multiple versions of a product (*profiles*), researchers are able to decompose the overall utilities of different product versions, and hence understand the role which each attribute plays in individuals' decision-making [58].

Applying the methodology of conjoint analysis to our context, we assume users view a third-party app as a combination of multiple app features. For example, if "information an app collects about a user's friends" constitutes an attribute of an app, the respective levels will be what or how much of friends' information is collected. Through analysis of how individuals evaluate versions of an app, we are able to infer how each factor, particularly revealing friends' personal information, affects a user's decision of adopting an app.

3.1.1 Determination of Apps' Attributes and Levels

Through conducting 18 semi-structured interviews with app users, we identified in our previous work [83] four attributes that are most frequently regarded as relevant to the choice of third-party apps. In addition, the interview results also helped to determine levels of these four app attributes. To allow for a direct comparison of results, we applied these insights also to the current context. In other words, we used the same app attributes and levels [83] which are summarized in Table 1.

3.1.2 Selection of Conjoint Analysis Method

There are two popular ways to conduct conjoint analyses: full-profile conjoint analysis and choice-based conjoint analysis. Typically, in a full-profile conjoint study, participants are asked to rank several versions of a product which differ regarding multiple attributes (see an example in Appendix A). Considering that each attribute has multiple levels, ranking even a small number of product versions represents a very high cognitive challenge to respondents [36]. Therefore, as is demonstrated in our previous research [81, 83], full-profile conjoint analysis studies include a non-trivial share of participant responses with low quality.

To address this problem, we decided to apply the methodology of choice-based conjoint analysis. In a choice-based conjoint study, respondents are asked to choose an alternative from a small set (normally 2 or 3) of profiles (*choice set*) [25] (see Figure 1). Participants then repeat this task for a limited number of choice sets, thereby providing adequate data for analysis. As a result, compared with full-profile conjoint analysis, the choice-based method poses less cognitive challenges to participants. We expect that by choosing this approach, we can obtain significantly higher quality responses.

3.1.3 Selection of App Profiles

We next discuss how to determine the number of choice sets to be included in the study. While there is no clear guidance on this issue, prior studies indicate that respondents are capable of managing 17 choice sets without problems [8], and a study on the commercial use of conjoint analysis reported a median value of 16 choice sets in typical conjoint study designs [113]. Based on these results, we included 16 choice sets in our study. Note here, in order to check for consistency of participants' responses, we set two choice sets to be the same. Therefore, our study included 15 distinct choice sets.

We adapted R code provided by Burda and Teuteberg [13] to create choice scenarios (choice sets) in our study. Specifically, with the help of the Algorithmic Experimental Design R package [110], we calculated a fractional factorial design from our full factorial design ($2 \times 2 \times 3 \times 3 = 36$ stimuli) by following a 5-step procedure described in [4]. Using this method, we derived a design including 15 different app profiles which were randomly combined to form the choice sets. In addition, in order to make the scenario more realistic, we also introduced the "no choice" option in each choice set. Therefore, we generated 15 different choice sets, with each of them including two app profiles and one "no choice" option.

3.1.4 Estimation of Conjoint Model

Hierarchical Bayes (HB) estimation takes into consideration that individuals have heterogeneous preferences for product-specific attributes and is generally preferred for analyzing choice-based conjoint models [88]. Without treating all individuals alike, the HB method allows not only for estimating a conjoint model on an aggregate level, but also for calculating parameter estimates associated with specific individuals, i.e., individual-level part-worth utilities. We further utilize the R package Bayesm [87] to conduct the HB estimation and to analyze our choice-based conjoint model.

3.2 Design of Survey Experiment

3.2.1 Treatments

Prior research indicates that individuals behave differently when anonymity is preserved than under circumstances with full information; in this case, identifiability and observability. We reviewed this literature in the related work section, but briefly summarize several results here. For example, by comparing results of F2F bargaining and anonymous bargaining in a classic behavioral experiment that aims to understand how agents cooperate with each other, Radner and Schotter find that F2F bargaining captured a higher percentage of gains from trade than anonymous bargaining [85]. Similarly, in another pair of comparative experiments, Roth and Malouf observe fewer equal splits and more disagreements in anonymous bargaining than in the F2F setting [90]. Interpreting these results, Siegel and Fouraker acknowledge that small differences in the social environment (such as the provisioning of anonymous communications) might lead to a large divergence in behavior [93]. Therefore, they argue that social variables, in particular anonymity, should either be systematically studied or controlled in behavioral experiments [93].

Applied to our context, we conjecture that anonymity plays a significant role in individuals' valuations of interdependent privacy. More specifically, we argue that app users will valuate their friends' information comparatively lower when they believe sharing friends' information with apps is anonymous compared to a full information scenario with observability of actions. In order to empirically investigate such an effect, we introduced the following 2 treatment scenarios regarding *sharing anonymity*:

1. Friends will <u>not</u> be able to discover who releases their information to apps (anonymous sharing).

2. Friends will be able to discover who releases their information to apps (identifiable sharing).

In addition, previous studies reveal that individuals' privacy con-

Attributes	Attribute Descriptions	Attribute Levels
Price	Price of the app	**$0.00**: The app is free **$1.99**: The app costs $1.99
Network Popularity	Percentage of a user's friends who installed the app	**5%**: 5% of a user's friends have installed the app **25%**: 25% of a user's friends have installed the app
Own Privacy	Information the app collects about a user	**None**: The app does not collect any information about a user **Basic profile**: The app collects a user's name, profile picture, gender, user ID, and any other information the user made public on his/her profile **Full profile**: The app collects a user's *Basic profile*, and in addition the user's valuable information, such as email address, birthday, photos, and location information
Friends' Privacy	Information the app collects about a user's friends	**None**: The app does not collect any information about a user's friends **Basic profile**: The app collects a user's friends' names, profile pictures, genders, user IDs, and any other information friends made public on their profiles **Full profile**: The app collects a user's friends' *Basic Profiles*, and in addition friends' valuable information, such as email addresses, birthdays, photos, and location information

Table 1: Summary of attributes and levels

cerns are influenced by whether or not information requests are context-relevant [68]. For example, Wang et al. [108] discover that while app users are typically unconcerned about giving away birthday information to a calendar app, they become uncomfortable when the app wants to collect information that is unrelated to the app's stated purpose. Motivated by the theory of contextual integrity [68] and the aforementioned empirical results, we also aim to explore how app data collection context impacts the value which app users place on their friends' information.

To this end, similar to how we account for sharing anonymity, we introduced the following 2 treatment scenarios regarding *context relevance*:

1. The information the app collects about user's friends is not useful for app's functionality (irrelevant context).

2. The information the app collects about user's friends is useful for app's functionality (relevant context).

To sum up, we included a total of 4 treatment conditions (2 sharing anonymity $\times$ 2 context relevance) in our study. We then randomly placed participants in one of the 4 treatments, which was then introduced as part of the task instructions. In addition, we displayed a short version of the instructions with the treatment conditions above each choice-based conjoint analysis question.

3.2.2 Procedure

After consenting to take part in the study, participants were randomly placed into one of the 4 treatments, and were provided with task instructions, where we offered definitions of app attributes and their corresponding levels. Next, they were presented with 16 questions (see Figure 1 for the app choice interface), which corresponded to the 16 choice sets in the conjoint analysis study. In each question, they were required to select their favorite alternative from two app versions and a "no choice" option. When participants selected a "no choice" option, it indicated that neither of the two provided app versions were preferred by them. Note here, in order to ensure that definitions of app attributes and levels were well conveyed to participants, we allowed them to revisit the instruction page during each app choice task.

After participants finished all 16 questions regarding their preferred choice of app profiles, they were asked to answer several demographic questions. In addition, since our paper aims not only to quantify the value of interdependent privacy and its dependency on sharing anonymity, but also to build a model to explain app users' privacy evaluation process, we also measured perceptual variables regarding users' privacy related attributes, beliefs and experiences (see details in later sections).

3.2.3 Recruitment and Ethical Considerations

We recruited participants from Amazon Mechanical Turk, a recruitment source that is popular for conducting online user experiments [35]. We restricted participation to Turkers who had completed over 50 Human Intelligence Tasks (HITs) with a HIT approval rating of 97% or better, as well as those who had United States IP addresses. In addition, eligible participants should have previously installed at least one app on their social network sites, so that they were familiar with the scenario setting of our study. We paid $1.50 to each participant after they completed the task.

Our study followed a protocol reviewed and approved by the IRB of the Pennsylvania State University. In addition, our survey-based investigation did not raise any significant ethical issues since it was a standard consumer study with an established study methodology and hypothetical choice situations, and it did not involve any deception.

3.3 Results of Choice-based Conjoint Study

3.3.1 Participant Data

Our survey study was conducted in September 2016. We collected a total of 1007 responses. After filtering out data based on conditions such as whether participants are US citizens, whether responses pass the check conditions implemented in the survey, and whether responses result in privacy values that are not outliers[3], our final sample included responses of 931 participants for data analysis. By comparing percentages of low quality responses between the current study and our prior work [83], we believe that

[3]For some responses, utilities associated with "$1.99" and "$0.00" are identical or nearly identical, which indicates zero or approaching zero utility change associated with per-dollar change. In this case, dollar equivalents for level changes in other attributes are either not determinable or abnormally large. Therefore, we did not include such responses in our analysis.

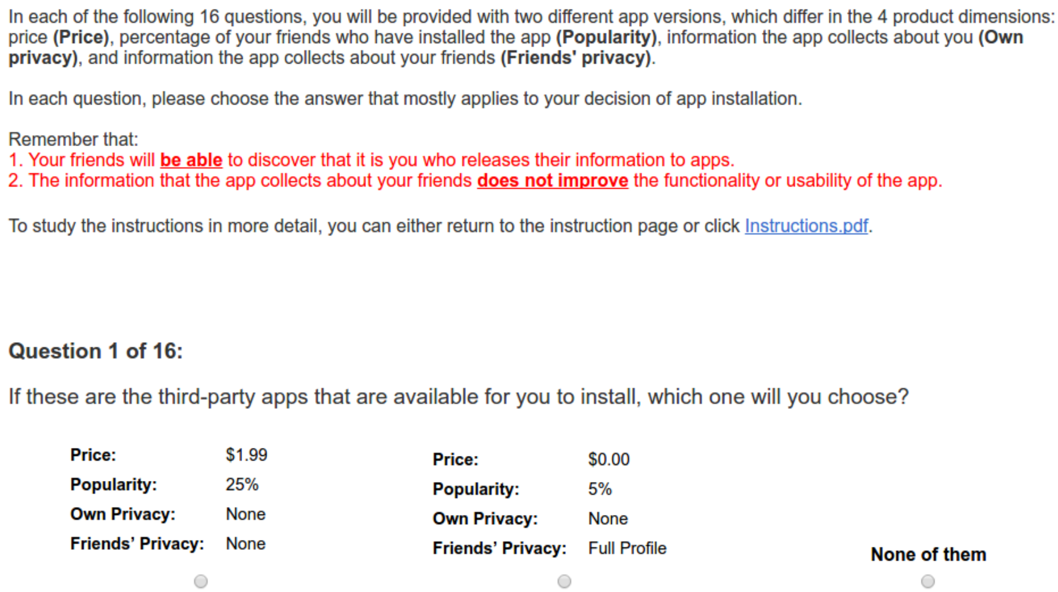


Figure 1: Screenshot of app choice interface

the choice-based conjoint method in our study has improved data quality by about 20%.

Of the 931 participants, 47.6% are male and 52.4% are female. In addition, our sample covers a wide range of age categories and education levels, ranging from 18 to over 50, and ranging from less than high school to higher education degrees such as master and PhD, respectively. In terms of income level, our participants have yearly incomes that range from less than $25,000 to more than $100,000, with a majority of them falling into the categories below $50,000.

Among the 931 participants, 234 were assigned to T1 (anonymous sharing & irrelevant context), 230 were assigned to T2 (identifiable sharing & irrelevant context), 239 were assigned to T3 (anonymous sharing & relevant context), and the remaining 228 were assigned to T4 (identifiable sharing & relevant context). Chi-square tests indicate that these four sample groups do not significantly differ regarding the demographic measures described above.

3.3.2 *Estimations of Privacy Values*

In this section, we first describe goodness-of-fit of the estimated conjoint model. Then, we show how to use the estimated model parameters to quantify privacy valuations. Note here, following the practice in Burda and Teuteberg [13], we did not use "no choice" data, i.e., "None of them" responses during the app choice tasks, to analyze the model.

We conducted two tests to assess goodness-of-fit of the estimated model. A likelihood ratio (LR) test was first performed to measure how well the model and its estimated parameters perform compared with having no model [104]. The test indicated that all the four estimated models (one model for each treatment condition) are statistically valid ($p < 0.001$ for all models), i.e., the null hypothesis that the estimated model and zero model are equal can be rejected. In addition, to assess the validity of our model, we calculated the hit rate by identifying the alternative with the highest probability in all 15 choice sets for each participant [104]. Each of the four models has a hit rate of more than 90%, demonstrating all these four models are well-fitted.

Next, we calculated dollar values for privacy following the approach described by Krasnova et al. [62]. Conjoint analysis allows us to calculate individual and aggregated part-worths (utilities), which denote the attractiveness of a specific attribute level. Based on the part-worths, we calculated utility changes between various attribute levels as well as corresponding dollar values for each attribute level change (see details in Appendix B). We show these results in Table 2, where the "Utility Change" column indicates aggregated utility changes, while the "Dollar Values" column displays averages of dollar values perceived by individuals.

From Table 2, we can access the dollar values which individuals place on different dimensions of own information and of friends' information. For example, we notice that in the scenario where sharing friends' information is anonymous and where such information is irrelevant to app's functionality (T1), individuals value their friends' *basic* information (corresponding to friends' privacy level change from "None" to "Basic profile) at $0.55, friends' *valuable* information (referring to friends' privacy level change from "Basic profile" to "Full profile") at $2.36, and friends' *full* profile information (matching friends' privacy level change from "None" to "Full profile") at $3.33.

We also observed from Table 2 that in most cases, dollar values which individuals place on their friends' information are slightly larger compared to their valuation of their own information. At the first glance, this observation might be counter-intuitive. However, friends' privacy value reported here is the dollar value that an individual places on the information of *all* of his/her friends.[4] Considering that our participants self-reported to have on average 263 friends on their preferred SNS, this means that the value for a

[4]We made it clear to participants that an app might collect information of all their friends by not only using and highlighting the words "friends", but also explicitly asking for the number of their social network friends.

Attributes	Level Change	Utility Change				Dollar Value			
		T1	T2	T3	T4	T1	T2	T3	T4
Price	\$0.00 ⇒ \$1.99	-3.43	-2.62	-3.60	-3.36	-1.99	-1.99	-1.99	-1.99
Own Privacy	None ⇒ Basic profile	-0.35	0.24	0.37	0.53	0.02	-1.38	0.20	0.34
	Basic profile ⇒ Full profile	-3.69	-2.73	-2.51	-2.80	-2.80	-2.28	-2.27	-2.36
	None ⇒ Full profile	-4.04	-2.48	-2.14	-2.27	-2.78	-3.66	-2.07	-2.02
Friends' Privacy	None ⇒ Basic profile	-0.60	-1.31	-0.17	-1.39	-0.55	-1.74	-0.02	-0.80
	Basic profile ⇒ Full profile	-3.37	-2.33	-1.82	-2.85	-2.36	-3.20	-1.49	-2.26
	None ⇒ Full profile	-3.97	-3.64	-1.99	-4.25	-3.33	-5.40	-2.09	-2.82

Table 2: Utility change and monetary value of change

single friend's personal information is very small (as small as a few cents) suggesting that individuals are *privacy egoists*.

3.3.3 *Effects of Sharing Anonymity and Context Relevance on Privacy Valuation*

We conducted a two-way analysis of variance (ANOVA) to investigate both the effects of sharing anonymity and context relevance on personal privacy valuation and interdependent privacy valuation.

Our analysis shows a significant main effect of sharing anonymity on the valuation of friends' basic information ($F(1,931) = 11.95$, $p = 0.001$), friends' valuable information ($F(1,931) = 6.33, p = 0.012$), and friends' full profile information ($F(1,931) = 5.03, p = 0.025$). More specifically, when sharing friends' information is anonymous, individuals value their friends' privacy significantly less than in the scenario where such sharing behavior is identifiable (see Figure 2, Figure 3, and Figure 4).

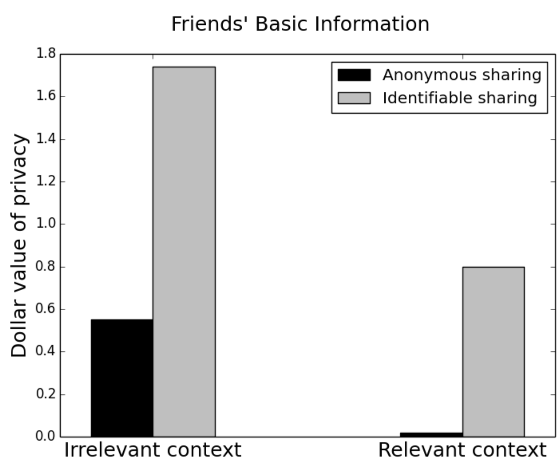


Figure 2: Effects of sharing anonymity and context relevance on valuation of friends' basic information

When it comes to the valuation of own personal privacy, we fail to detect a significant impact of sharing anonymity. In other words, the condition as to whether or not sharing friends' information is anonymous does not affect how individuals value their own basic information ($F(1,931) = 1.72, p = 0.189$), own valuable information ($F(1,931) = 0.14, p = 0.708$), or own full profile information ($F(1,931) = 0.90, p = 0.344$).

As to the condition of context relevance, we find it to significantly affect valuation of interdependent privacy. Individuals place lower values on friends' basic information ($F(1,931) = 6.61, p = 0.010$), valuable information ($F(1,931) = 7.92, p = 0.005$), and full profile

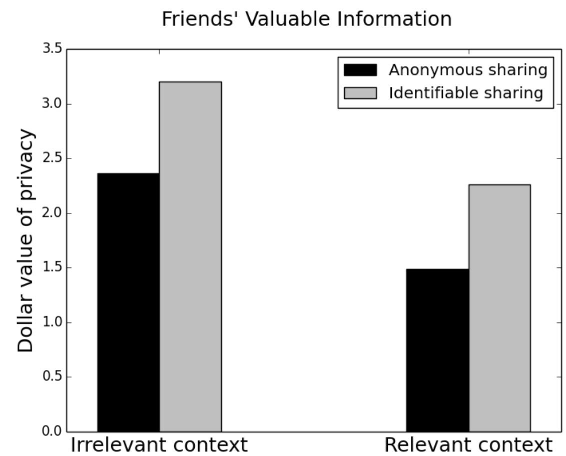


Figure 3: Effects of sharing anonymity and context relevance on valuation of friends' valuable information

information ($F(1,931) = 9.32, p = 0.002$), when they believe that information improves apps' functionality compared to the alternative scenario (see Figure 2, Figure 3, and Figure 4).

The impact of context relevance regarding the valuation of own valuable data is insignificant ($F(1,931) = 0.15, p = 0.696$); however, we observe that the treatment effect is significant for the value which individuals place on their own basic information ($F(1,931) = 3.86, p = 0.050$) and full profile information ($F(1,931) = 7.17, p = 0.008$). This might indicate that the condition of context relevance, even though only information is given to the individual about the relevance of app's usage of *friends'* personal information, has a partial spillover effect on the valuation of their *own* privacy.

In addition, we tested for any possible interactions between sharing anonymity and context relevance on privacy valuation. However, such effects do not exist for either own privacy valuation or interdependent privacy valuation.

4. INVESTIGATION OF DETERMINANTS OF PRIVACY VALUE WITH SEM

By applying choice-based conjoint analysis, we quantified the dollar values which app users place on their own and friends' privacy. We next aim to position the conjoint study results in an SEM model to investigate what drives the valuation of personal and interdependent privacy. More specifically, we seek to understand how factors such as different dimensions of privacy concerns, their antecedents, sharing anonymity, as well as context relevance affect the valuations of app users' own and their friends' information.

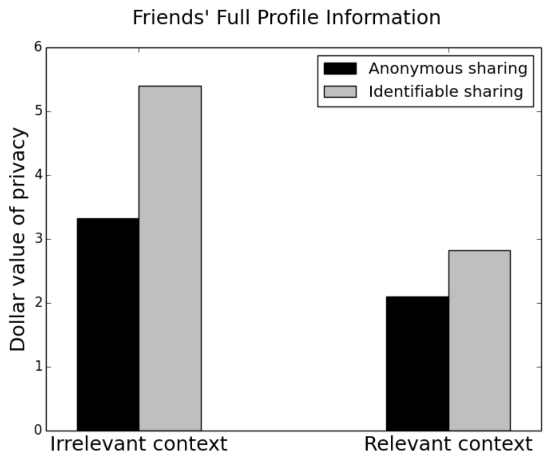


Figure 4: Effects of sharing anonymity and context relevance on valuation of friends' full profile information

In this section, we first identify factors that affect users' valuations of privacy based on the existing literature. We next build an SEM model to examine relationships between these identified factors and the measured privacy valuations.

4.1 Hypotheses and Research Model

When individuals provide their information to other parties, a *social contract*, which is generally understood as the expectation that these parties will manage personal information properly [77], is formed [16]. If individuals believe their personal information has been misused, they may consider such an implied contract breached [77, 23], and hence lower their trust assessment associated with the involved parties. In addition, prior research shows that in the electronic commerce context, an online consumer's privacy being intruded by a single online company could lead to the perception of information misuse by the entire community of online sellers [75]. In particular, individuals who have been exposed to or have been the victim of personal information abuses could be more aware of what actions other parties could take to invade privacy [2]. Such awareness might in turn lead to the reduction of their trust in online companies. Applying this to our context, we argue that the more past privacy invasion experiences individuals have, the less likely they are to trust apps' practices to protect their privacy. Therefore, we hypothesize:

Hypothesis 1: *There is a negative relationship between individuals' past privacy invasion experiences and their trust in apps' data practices.*

Previous studies demonstrate trust can enhance the evaluation of benefits, and can mitigate privacy concerns [74]. In particular, Hoffman et al. [47] argue that in the setting of online commerce, trust creates positive attitudes toward Web retailers. More specifically, trust refers to individuals' feelings that they will gain the benefits they expect without suffering negative consequences [74]. In this manner, we believe that app users who trust apps' data practices are less likely to be concerned when releasing their own personal information to apps. Hence, we making the following hypothesis:

Hypothesis 2: *There is a negative association between individuals' trust in apps' practices and their concerns for own information privacy.*

Disposition to value privacy is a personality attribute reflecting an individual's inherent need (or general tendencies) to manage personal information space [115]. Therefore, as opposed to individuals who tend to be more open regarding the sharing of their personal information, individuals with a higher disposition to value privacy will also express a higher level of concern when disclosing their own personal information to others. Hence, we argue:

Hypothesis 3: *Individuals' dispositions to value privacy are positively related to their concerns for own information privacy*.

Empirical evidence in numerous studies reveals that control is one of the key factors that affects privacy concerns [27, 77]. For example, it has been found that individuals' perceptions of control over dissemination of personal information are negatively related to privacy concerns [66, 114]. Additionally, research has provided evidence that, in general, individuals will have fewer privacy concerns when they believe they can control the release and dissemination of their personal information [100, 66]. To confirm these findings, we also make the following hypothesis:

Hypothesis 4: *Individuals' perceived privacy control is negatively associated with their concerns for own information privacy.*

Prior research shows that receiving negative news reports regarding privacy, such as stories about the gathering and misusing of personal information, contributes to individuals' privacy concerns [70]. Therefore, we argue that the more knowledge about privacy an individual has, the higher the level of concerns he/she will express over both own and friends' privacy. Hence, we hypothesize:

Hypothesis 5: *Privacy knowledge is positively related to individuals' concerns for their own information privacy.*

Hypothesis 6: *Privacy knowledge is positively related to individuals' concerns for their friends' information privacy.*

Experimental results provide substantial evidence for the existence of *other-regarding preferences* [22, 98]. In a nutshell, the theory of other-regarding preferences indicates that individuals are not purely selfish, but rather care about others' well being. However, they differ in the extent to which they care about others, which can be determined by measuring the strength of their other-regarding preferences. Applying this theory to our context, we believe individuals who have higher other-regarding preferences express higher levels of privacy concerns over their friends' information. Hence, we argue:

Hypothesis 7: *Individuals' other-regarding preferences are positively related to their concerns for friends' information privacy.*

In addition, it is reasonable to assume that while keeping other factors constant, more privacy-concerned individuals exhibit higher privacy valuations (as measured in the conjoint study). It follows that we hypothesize:

Hypothesis 8: *Individuals' concerns for their own information privacy are positively associated with the perceived monetary value of their own information.*

Hypothesis 9: *Individuals' concerns for friends' information privacy are positively associated with the perceived monetary value of their friends' information.*

Recall that in the conjoint analysis study, we introduced four treatment conditions manipulating whether disclosure of friends' information is anonymous (sharing anonymity), and whether friends' information is necessary for apps' functionality (context relevance).

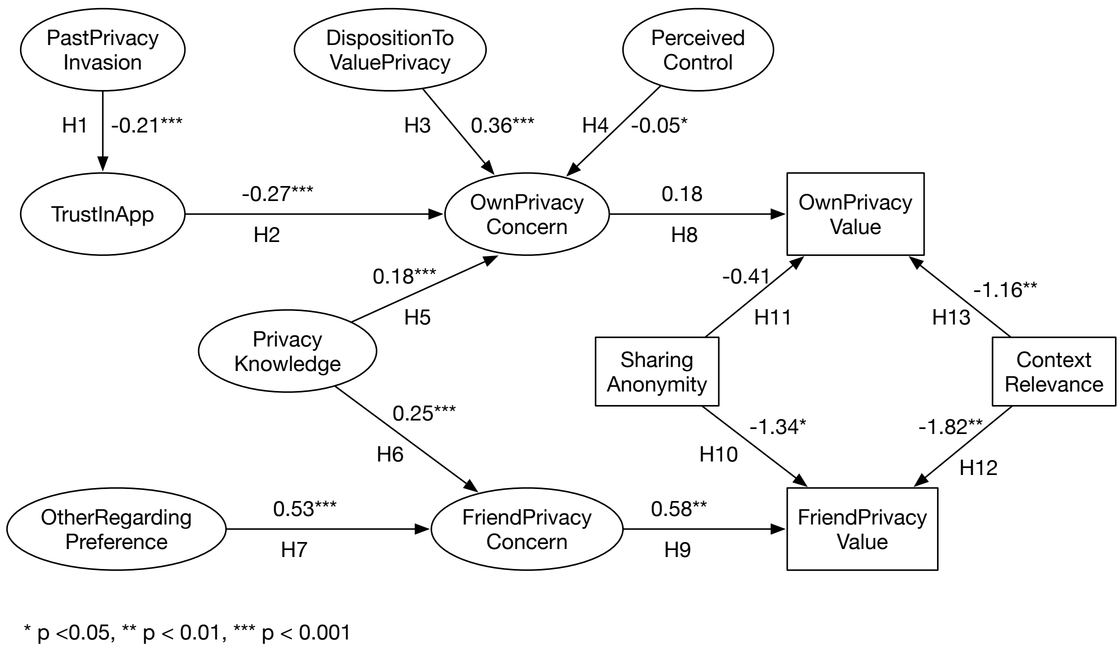


Figure 5: SEM explaining privacy valuation

The conjoint analysis results show that both sharing anonymity and context relevance impact the value which app users place on their friends' information. When it comes to the valuation of own information, we detect a (partial) significant spill-over effect of context relevance, but not for sharing anonymity.

Here, we integrate these effects in the SEM model not only for building a more comprehensive model of privacy valuation, but also for providing an additional method to examine significances of these effects. Therefore, we assume:

Hypothesis 10: *Under the condition of anonymous sharing, individuals place lower monetary values on their friends' information compared with identified sharing.*

Hypothesis 11: *Under the condition of anonymous sharing, individuals place lower monetary values on their own information compared with identified sharing.*

Hypothesis 12: *Under the condition of context-relevant data collection, individuals place lower monetary values on their friends' information compared with context-irrelevant data collection.*

Hypothesis 13: *Under the condition of context-relevant data collection, individuals place lower monetary values on their own information compared with context-irrelevant data collection.*

We present the research model, which is based on H1 $\sim$ H13, in Figure 5.

4.1.1 Measurement Scale Development

To the extent possible, we adapted measurement scales for the main constructs in this study from prior research to fit the app adoption context.

Adapting from Smith et al. [95], 4 questions were used to assess past privacy invasion experiences. Trust in apps was measured by a shortened 4-item version of trust measures from Fogel and Nehmad [33], Krasnova and Veltri [63], and Dwyer et al. [29]. To measure privacy knowledge, we used 4 items derived from Park et al. [73]. Disposition to value privacy and perceived control were measured based on the 3-item scale and 4-item scale developed in [115], respectively. With respect to other-regarding preferences, we applied 5 items modified from the actively caring scale in [86]. When it comes to privacy concern, four items derived from [95] are used to assess privacy concerns for one's own information. A similar set of 4 questions, which was also derived from [95], was applied to measure individuals' concern for friends' privacy. All items were measured on a Likert-type scale with 1 = strongly disagree to 5 = strongly agree. The exact questions are provided in Appendix C.

In the conjoint analysis study, we measured three dimensions of privacy value: value of basic information, value of valuable information, and value of full profile information. Since the full profile information includes both basic and valuable information, we limit our model to the study of valuation of full profile information. As such, we used the monetary value that individuals place on their own full profile information to represent own privacy valuation in the SEM model. Similarly, the valuation of friends' privacy in the model is represented by the dollar value of friends' full profile information.

In addition, sharing anonymity and context relevance in the SEM model correspond to the treatment scenarios we set in the conjoint analysis survey. For example, a value of 1 of sharing anonymity indicates that sharing friends' information cannot be identified, and a value of 1 of context relevance signifies that friends' information collected by an app improves apps' functionality.

4.1.2 Evaluation of the Measurement Model

The measurement model is evaluated in terms of both convergent validity and discriminant validity. Convergent validity measures the degree to which different attempts to measure the same construct agree [21]. Two tests are conducted to determine the convergent validity of the scales: Cronbach's alpha and composite reliability of constructs. We present the test results in Table 3. As

	Cronbach's Alpha	Composite Reliability	Trust InApp	Privacy Knowledge	PastPrivacy Invasion	Disposition ToValue Privacy	Other Regarding Preference	Perceived Control	Own Privacy Concern	Friend Privacy Concern
TrustInApp	0.88	0.89	0.81							
PrivacyKnowledge	0.88	0.88	-0.25	0.80						
PastPrivacyInvasion	0.78	0.78	-0.20	0.10	0.69					
DispositionToValuePrivacy	0.87	0.87	-0.14	0.08	0.24	0.84				
OtherRegardingPreference	0.71	0.74	-0.08	0.33	-0.02	0.08	0.61			
PerceivedControl	0.91	0.91	0.49	-0.28	-0.19	0.16	-0.06	0.84		
OwnPrivacyConcern	0.89	0.89	-0.41	0.25	0.21	0.60	0.20	-0.14	0.82	
FriendPrivacyConcern	0.93	0.93	-0.19	0.25	0.11	0.31	0.37	-0.10	0.58	0.88

Table 3: Evaluations of measurement model

is shown in Table 3, the Cronbach's alpha values for all scales are larger than 0.7; an indication of adequateness proposed by Nunnally [71]. In addition, composite reliabilities of our constructs exceed Nunnally's [71] criterion of 0.7. Both of these tests support the convergent validity of our measurement model.

Discriminant validity evaluates to which degree measures of different constructs are distinct from each other [14]. Discriminant validity is achieved when the square root of the variance shared between a construct and its measures is greater than the correlations between the construct and any other constructs in the model. We show the results in Table 3. We observe from Table 3 that the correlations among constructs, i.e., non-diagonal elements, are less than the square roots of shared variance, i.e., diagonal elements, indicating our model fulfills the requirements of discriminant validity.

In addition, we also conduct confirmatory factor analysis to provide an additional method to assess our measurement model. Specifically, Mean Square Error of Approximation (RMSEA) value and Comparative Fit Index (CFI) are used here. A RMSEA value of 0.06 or less, or a CFI value of 0.90 or greater indicates the model fit is acceptable [53]. Our measurement model has $RMSEA = 0.04$ and $CFI = 0.93$; further indicating that our measurement model is of high quality.

4.1.3 Evaluation of the Path Model

We first discuss goodness-of-fit data of the model. In SEM, the chi-square test is a frequently reported descriptive measure of fit. Usually, a chi-square test with a $p-$value exceeding 0.05 demonstrates a model is a good fit (i.e., significance might indicate a bad fit) [6]. Due to chi-square tests' sensitiveness to sample size, other goodness-of-fit criteria, i.e., RMSEA and CFI, are also used [50].

The goodness-of-fit data of our model is $\chi^2(579) = 1841.89$, $p = 0.00; RMSEA = 0.05$; and $CFI = 0.93$. Despite the significant result of the chi-square test, which is sensitive to sample size, the RMSEA value and CFI together indicate that our model fit is acceptable.

We next test H1 $\sim$ H13, which should be evaluated based on the sign and statistical significance (assessed by z-test) for corresponding paths in the model. We show the test results in Figure 5.

Our results support most of the associations we hypothesized. Individuals' past privacy invasion experiences are found to be significantly and negatively associated with their trust in apps' data practices (H1 is supported), which in turn has a significant and negative impact on concerns for own personal privacy (H2 is supported). In support of H3 and H4, the positive relationship between individuals' disposition to value privacy and concerns for own privacy, and the negative association between individuals' perceived control and own privacy concerns, are both found to be significant. In addition, individuals' privacy knowledge is found to significantly impact concerns for both personal and friends' information privacy (H5 and H6 are supported). Further, the proposed impact of other-regarding preferences on concerns towards friends' information privacy is also significant (H7 is supported).

When it comes to the relationship between privacy concerns and monetary value of personal privacy, we do not find such an association which is statistically significant (H8 is not supported). In contrast, we observe a significant effect explaining the relationship between concerns for others' privacy and the valuation of friends' information (H9 is supported).

H10 $\sim$ H13 postulate the impacts of treatment conditions (sharing anonymity and context relevance) on privacy valuation. In support of H12 and H13, the negative impact of context relevance on both own privacy valuation and valuation towards friends' information are found to be significant. In addition, sharing anonymity is also significantly and negatively associated with the value which individuals place on their friends' privacy (H10 is supported). However, the proposed negative impact of sharing anonymity on how app users valuate their own personal information is insignificant (H11 is not supported). These results are in line with the findings we have discussed earlier in the conjoint analysis study.

4.1.4 Discussion of SEM Results

Through conducting an SEM analysis, we explore factors that drive the valuations of own privacy and interdependent privacy. In particular, we examine how conditions such as sharing anonymity and context relevance affect privacy valuations.

Our results suggest that individuals' interdependent privacy valuations are partly determined by their personal attributes and experiences, which is similar to findings in [67]. For example, through raising privacy concerns for friends' information, app users' inherent other-regarding preferences play an important role in shaping how they value others' privacy. Similarly, through the mediation of concerns towards friends' privacy, privacy knowledge impacts the values which app users place on friends' information. This indicates that educating app users about practices impacting interdependent privacy might be a viable way to increase their valuation of interdependent privacy.

Our results further demonstrate that individuals' valuations of their friends' privacy can also be influenced by environmental settings. In particular, the value of interdependent privacy is found to be sensitive to the treatment regarding anonymity. It appears that when individuals believe their actions of disclosing friends' information to apps can be identified, they will think twice before taking such actions. Similarly, when friends' information collected by apps is irrelevant to apps' stated purposes, individuals will be more reluctant to share such information. Therefore, besides raising individuals' interdependent privacy concerns, an alternative way to protect those who might suffer from interdependent privacy is to manipulate exogenous conditions, e.g., by making the sharing of friends' data identifiable or by informing app users whether data collection is context relevant.

Similar to their concerns about friends' privacy, users' concerns towards their own privacy is affected by their personal beliefs and experiences. In particular, we find individuals' inherent needs to manage personal information space, and beliefs regarding whether or not they are able to control privacy influence how concerned they are about their personal privacy.

When it comes to users' valuation of personal privacy, our results suggest that the condition as to whether friends' information collected by an app is relevant to the app's functionality also has a significant impact. Given that context relevance does not differ in terms of apps' practices of accessing users' own personal information (as per the experimental instructions), this suggests a spillover effect [26, 91]. In other words, individuals might believe that their own information also contributes to app's functionality when they know this is the case for friends' information.

Although the empirical results provide overall support for the research model, they also reveal a few unexpected relationships that are inconsistent with what we have hypothesized. Specifically, the proposed positive associations between privacy concern for personal information and the perceived value of such information is not confirmed. This seemingly counter-intuitive result might be attributed to the nature of conjoint analysis. As discussed earlier, conjoint analysis is a method to uncover the hidden rules individuals apply to make trade-off decisions over different attributes. Applied to our context, the results we derive from conjoint analysis study are reflections of trade-offs participants make among app attributes, which include both personal privacy and friends' privacy. One thing to note here is that in the conjoint analysis survey, we highlighted treatment scenarios, i.e., 4 conditions regarding sharing anonymity and context relevance, not only during task instructions, but also at the beginning of each conjoint analysis question. Such emphasis might lead our participants to pay more attention to friends' privacy, and therefore may affect their valuations for own privacy. In this manner, even if users express high privacy concerns for their personal information, it does not necessarily correspond to equally high valuations for such information.

The insignificance of sharing anonymity in reducing users' perceived value of their own information makes sense since we would not expect a spillover effect in this case. As individuals in our study setup know that they are sharing their own information, the condition of sharing anonymity would not play a role in app users' valuation of their own privacy. (Of course, in practice users may not always pay attention to privacy conditions associated with an adoption decision, or may not fully understand these terms as they are often presented in user-*un*friendly ways [40].)

5. DISCUSSION

In this section, we present the emerging themes and practical design implications of our study. In addition, we also offer policy suggestions that are motivated by our research findings.

5.1 Implications for Privacy by Redesign

Our study contributes to a better understanding of individuals' perceptions, knowledge and preferences regarding interdependent privacy, thereby yielding implications for the "privacy by redesign" debate [17]:

5.1.1 Design to inform about data sharing anonymity

Our results highlight that informing individuals about whether or not sharing friends' information with apps is anonymous affects how they value interdependent privacy. Given that, a viable way to protect friends' privacy is not only to make such information sharing observable, but also to inform app users that the behavior of sharing friends' data is identifiable. For example, concrete mechanisms should be proposed so that when individuals share their friends' data, their friends will be notified about these sharing behaviors, or at least can access a permanent and easily accessible record of such actions (e.g., see the logging mechanism for mobile app behaviors proposed in Petracca et al. [76]). In addition, apps' authorization dialogues can be appropriately modified, so that they convey the information to app users that sharing friends' information will be later discoverable by friends.

We are cautiously optimistic that platform providers would be inclined to assist users in limiting the unwanted flow of information to an *outside* party, i.e., app developers. While a platform provider like Facebook benefits from business relationships with third-party developers (like Zynga), it should be cautious about bulk data transfers of their most valuable asset, i.e., user data. As the notification interfaces and authorization dialogues are provided by the platform, we see potential for improvements and limits to bulk data sharing.

5.1.2 Design to reflect data collection relevance

Research has proven that presenting privacy information in a clearer fashion to users, when they are making adoption decisions, can assist users in choosing less privacy-invasive apps [61, 107]. Our study demonstrates that data collection context affects how users value their friends' information. Therefore, in order to help app users make well-informed decisions, it would be useful to revise apps' privacy notice dialogues so that they explicitly inform users whether an app's practice of collecting data is necessary for an app's functionality. The input for such dialogues can stem from technical approaches which reverse-engineer apps to infer their usage of requested information [28, 32].

5.1.3 Design to control flow of friends' information

Our work indicates that app users are privacy egoists [81] not only in that they place on average less than a few cents on full profile information of a single friend, but also due to the fact that they are eager to reveal friends' data when they believe such disclosure behaviors result in better app performance. As such, relying on individuals themselves to protect friends' privacy is likely not adequate. Therefore, affected friends of app users should be involved more directly in the decision-making process. For example, designs that enable mutual agreements regarding sharing others' data, e.g., reciprocal designs that allow a user to share others' information if and only if he/she also lets others to share his/her information, should be implemented. Alternatively, we can also introduce easy-to-use mechanisms that empower affected individuals to unilaterally decide whether or not to allow their information to be shared by others.

5.2 Insights into Privacy Policy Discussions

Our study also contributes to policy discussions on app privacy, particularly on the problem of interdependent privacy.

5.2.1 Emphasize the role of government interventions

The central aspect of the problem of interdependent privacy is the existence of negative externalities, i.e., those who install apps that collect personal information of friends do not directly suffer from interdependent privacy harms. Similar to what economists generally suggest to deal with negative externalities [99], regulatory interventions should be put into place to deal with the problem space of interdependent privacy (including social app adoption scenarios). For example, policies or laws (e.g., privacy baseline regulations [106]) need to be introduced to rigorously limit apps' practice of collecting friends' data in bulk.

In addition, as aforementioned, it is not adequate to rely on app users to protect their friends' privacy since app users are often privacy egoists. This further emphasizes the importance of government intervention to address the issue of interdependent privacy.

5.2.2 *Promote education on privacy*

Our work confirms that privacy knowledge impacts the values which app users place on friends' information. This indicates that educating app users about practices impacting interdependent privacy might be a viable way to increase their valuation of interdependent privacy. Therefore, policy makers should consider introducing policies which integrate privacy in educational programs. We have previously tested the introduction of relatively advanced measurement methodologies for online tracking in the context of an educational program with overall positive results [69].

6. CONCLUSIONS

To the best of our knowledge, this paper is one of the first formal studies to investigate the impact of anonymity on privacy decision-making and, in particular, on the valuation of interdependent privacy. Through conducting a choice-based conjoint analysis study with different treatment scenarios, we quantify the economic value app users place on both their own and friends' information, and also examine the impact of treatment conditions on privacy valuation. We also built and estimated an SEM model to explore how factors such as individuals' personal beliefs, attributes, experiences, as well as environmental factors, i.e., sharing anonymity and context relevance, contribute to individuals' perceived value of both personal and friends' privacy. Our research findings yield valuable insights, such as implications for the redesign of apps' privacy notice and permission dialogues, as well as suggestions to introduce new privacy polices, for better addressing individuals' own and their friends' privacy preferences.

Several limitations should be considered. In the conjoint analysis survey, we make the treatment scenarios salient by not only emphasizing them during task instructions, but also highlighting them in each conjoint choice question. Given that these treatment scenarios are highly related to the collection of friends' information, this implementation may give additional emphasis to the importance of interdependent privacy, and thereby reduce the perceived importance of personal privacy. Therefore, the low valuations for the data of individual friends are particularly notable. Nevertheless, one should proceed with care when comparing the absolute values for personal privacy and friends' privacy, and we recommend to use the results across the slightly different settings and methodologies in our related works as a joint basis for evaluations [81, 83, 82].

In addition, in our conjoint analysis study, the choice of an app is still at a hypothetical level, where participants did not really "gamble with their own money". Therefore, compared with real world scenarios where real costs can be incurred, the monetary value which participants put on others' data might be overestimated in the current study. However, given that even in the hypothetical scenario where individuals could show themselves from a desirable side at no cost, they prefer to disclose others' data when the scenario states they will not suffer directly from such behaviors, the problem of interdependent privacy may stand out even more in real world situations.

Further, we restrict the investigation of interdependent privacy valuation to the scenario of app adoption. However, other contexts, such as data analytics [19], location privacy [72], and genetic privacy [55, 109], also emphasize the issue of interdependent privacy. Therefore, it is prudent to also study interdependent privacy valuation in these settings in order to contribute to the generalizability of our findings.

Acknowledgments: We thank S. Shyam Sundar for insightful discussions during the design stage for this research study. We also thank Mary Beth Rosson, Peng Liu and the anonymous reviewers for their detailed suggestions for improvements of the manuscript. We further acknowledge the comments we received for our presentation at the Federal Trade Commission PrivacyCon 2017 conference. The research activities of Jens Grossklags are supported by the German Institute for Trust and Safety on the Internet (DIVSI).

7. REFERENCES

[1] A. Acquisti and J. Grossklags. Losses, gains, and hyperbolic discounting: An experimental approach to information security attitudes and behavior. In *Proceedings of the 2nd Annual Workshop on Economics and Information Security*, 2003.

[2] A. Acquisti and J. Grossklags. Privacy and rationality in individual decision making. *IEEE Security & Privacy*, 3(1):26–33, 2005.

[3] A. Acquisti and J. Grossklags. An online survey experiment on ambiguity and privacy. *Communications & Strategies*, 88(4):19–39, 2012.

[4] H. Aizaki and K. Nishimura. Design and analysis of choice experiments using R: A brief introduction. *Agricultural Information Research*, 17(2):86–94, 2008.

[5] F. Alpizar, F. Carlsson, and O. Johansson-Stenman. Anonymity, reciprocity, and conformity: Evidence from voluntary contributions to a national park in Costa Rica. *Journal of Public Economics*, 92(5):1047–1060, 2008.

[6] P. Barrett. Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, 42(5):815–824, 2007.

[7] C. Bauer, J. Korunovska, and S. Spiekermann. On the value of information - What Facebook users are willing to pay. *Proceedings of the European Conference on Information Systems (ECIS)*, 2012.

[8] M. Bech, T. Kjaer, and J. Lauridsen. Does the number of choice sets matter? Results from a web survey applying a discrete choice experiment. *Health Economics*, 20(3):273–286, 2011.

[9] A. Beresford, D. Kübler, and S. Preibusch. Unwillingness to pay for privacy: A field experiment. *Economics Letters*, 117(1):25–27, 2012.

[10] A. Besmer and H. Richter Lipford. Moving beyond untagging: Photo privacy in a tagged world. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1563–1572, 2010.

[11] G. Biczók and P. Chia. Interdependent privacy: Let me share your data. In A.-R. Sadeghi, editor, *Financial Cryptography and Data Security*, volume 7859 of *Lecture Notes in Computer Science*, pages 338–353. Springer, 2013.

[12] R. Böhme and J. Grossklags. Trading agent kills market

information: Evidence from online social lending. In *Proceedings of the 9th Conference on Web and Internet Economics (WINE)*, pages 68–81, 2013.

[13] D. Burda and F. Teuteberg. Understanding the benefit structure of cloud storage as a means of personal archiving – A choice-based conjoint analysis. In *Proceedings of the European Conference on Information Systems (ECIS)*, 2014.

[14] D. Campbell and D. Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81, 1959.

[15] B. Carminati and E. Ferrari. Collaborative access control in on-line social networks. In *2011 International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*, pages 231–240, 2011.

[16] E. Caudill and P. Murphy. Consumer online privacy: Legal and ethical issues. *Journal of Public Policy & Marketing*, 19(1):7–19, 2000.

[17] A. Cavoukian and M. Prosch. Privacy by redesign: Building a better legacy. *Information Privacy Commissioner Ontario*, pages 1–8, 2011.

[18] G. Charness and U. Gneezy. What's in a name? Anonymity and social distance in dictator and ultimatum games. *Journal of Economic Behavior & Organization*, 68(1):29–35, 2008.

[19] M. Chessa, J. Grossklags, and P. Loiseau. A game-theoretic study on non-monetary incentives in data analytics projects with privacy implications. In *Proceedings of the 2015 IEEE 28th Computer Security Foundations Symposium (CSF)*, pages 90–104, 2015.

[20] H. Cho and A. Filippova. Networked privacy management in Facebook: A mixed-methods and multinational study. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pages 503–514, 2016.

[21] T. Cook, D. Campbell, and A. Day. *Quasi-experimentation: Design & analysis issues for field settings*, volume 351. Houghton Mifflin, 1979.

[22] D. Cooper and J. Kagel. Other regarding preferences: A selective survey of experimental results. `http://myweb.fsu.edu/djcooper/research/otherregard.pdf`, forthcoming.

[23] M. Culnan. Consumer awareness of name removal procedures: Implications for direct marketing. *Journal of Direct Marketing*, 9(2):10–19, 1995.

[24] G. Danezis, S. Lewis, and R. Anderson. How much is location privacy worth? In *Proceedings of the Workshop on the Economics of Privacy (WEIS)*, 2005.

[25] W. S. DeSarbo, V. Ramaswamy, and S. H. Cohen. Market segmentation with choice-based conjoint analysis. *Marketing Letters*, 6(2):137–147, 1995.

[26] D. Dickinson and R. Oxoby. Cognitive dissonance, pessimism, and behavioral spillover effects. *Journal of Economic Psychology*, 32(3):295–306, 2011.

[27] T. Dinev and P. Hart. Internet privacy concerns and their antecedents-measurement validity and a regression model. *Behaviour & Information Technology*, 23(6):413–422, 2004.

[28] Q. Do, B. Martini, and K. Choo. Enhancing user privacy on Android mobile devices via permissions removal. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*, pages 5070–5079, 2014.

[29] C. Dwyer, S. Hiltz, and K. Passerini. Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace. In *Proceedings of the Americas Conference on Information Systems (AMCIS)*, 2007.

[30] M. Eastlick, S. Lotz, and P. Warrington. Understanding online b-to-c relationships: An integrated model of privacy concerns, trust, and commitment. *Journal of Business Research*, 59(8):877–886, 2006.

[31] S. Egelman, A. P. Felt, and D. Wagner. Choice architecture and smartphone privacy: There's a price for that. In *The Economics of Information Security and Privacy*, pages 211–236. Springer, 2013.

[32] W. Enck, P. Gilbert, S. Han, V. Tendulkar, B. Chun, L. Cox, J. Jung, P. McDaniel, and A. Sheth. TaintDroid: An information-flow tracking system for realtime privacy monitoring on smartphones. *ACM Transactions on Computer Systems*, 32(2):5:1–5:29, 2014.

[33] J. Fogel and E. Nehmad. Internet social network communities: Risk taking, trust, and privacy concerns. *Computers in Human Behavior*, 25(1):153–160, 2009.

[34] N. Good, J. Grossklags, D. Thaw, A. Perzanowski, D. Mulligan, and J. Konstan. User choices and regret: Understanding users' decision process about consensually acquired spyware. *I/S: A Journal of Law and Policy for the Information Society*, 2(2):283–981, 2006.

[35] J. Goodman, C. Cryder, and A. Cheema. Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, 26(3):213–224, 2013.

[36] P. Green and V. Srinivasan. Conjoint analysis in consumer research: Issues and outlook. *Journal of Consumer Research*, 5(2):103–123, 1978.

[37] P. Green and V. Srinivasan. Conjoint analysis in marketing: New developments with implications for research and practice. *The Journal of Marketing*, 54(4):3–19, 1990.

[38] J. Grossklags and A. Acquisti. When 25 cents is too much: An experiment on willingness-to-sell and willingness-to-protect personal information. In *Proceedings of the Workshop on the Economics of Information Security (WEIS)*, 2007.

[39] J. Grossklags and N. Barradale. Social status and the demand for security and privacy. In E. De Cristofaro and S. Murdoch, editors, *Privacy Enhancing Technologies*, volume 8555 of *Lecture Notes in Computer Science*, pages 83–101. Springer, 2014.

[40] J. Grossklags and N. Good. Empirical studies on software notices to inform policy makers and usability designers. In S. Dietrich and R. Dhamija, editors, *Financial Cryptography and Data Security: 11th International Conference, FC 2007, and 1st International Workshop on Usable Security, USEC 2007*, pages 341–355. Springer, 2007.

[41] S. Gürses, C. Troncoso, and C. Diaz. Engineering privacy by design. In *Proceedings of the Conference on Computers, Privacy & Data Protection*, 2011.

[42] W. Güth, R. Schmittberger, and B. Schwarze. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4):367–388, 1982.

[43] I.-H. Hann, K.-L. Hui, S.-Y. T. Lee, and I. Png. Overcoming online information privacy concerns: An information-processing theory approach. *Journal of*

Management Information Systems, 24(2):13–42, 2007.

[44] I.-H. Hann, K.-L. Hui, T. Lee, and I. Png. Online information privacy: Measuring the cost-benefit trade-off. In *Proceedings of the International Conference on Information Systems (ICIS)*, 2002.

[45] H. Harkous and K. Aberer. "If you can't beat them, join them": A usability approach to interdependent privacy in cloud apps. In *Proceedings of the Seventh ACM on Conference on Data and Application Security and Privacy*, pages 127–138, 2017.

[46] H. Harkous, R. Rahman, B. Karlas, and K. Aberer. The curious case of the PDF converter that likes Mozart: Dissecting and mitigating the privacy risk of personal cloud apps. *Proceedings on Privacy Enhancing Technologies*, 2016(4):123–143, 2016.

[47] D. Hoffman, T. Novak, and M. Peralta. Building consumer trust online. *Communications of the ACM*, 42(4):80–85, 1999.

[48] E. Hoffman, K. McCabe, K. Shachat, and V. Smith. Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, 7(3):346–380, 1994.

[49] E. Hollenbaugh and M. Everett. The effects of anonymity on self-disclosure in blogs: An application of the online disinhibition effect. *Journal of Computer-Mediated Communication*, 18(3):283–302, 2013.

[50] D. Hooper, J. Coughlan, and M. Mullen. Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, 6(1):53–60, 2008.

[51] H. Hu, G.-J. Ahn, and J. Jorgensen. Multiparty access control for online social networks: Model and mechanisms. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1614–1627, 2013.

[52] H. Hu, G.-J. Ahn, Z. Zhao, and D. Yang. Game theoretic analysis of multiparty access control in online social networks. In *Proceedings of the 19th ACM Symposium on Access Control Models and Technologies*, pages 93–102, 2014.

[53] L. Hu and P. Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1):1–55, 1999.

[54] B. Huberman, E. Adar, and L. Fine. Valuating privacy. *IEEE Security & Privacy*, 3(5):22–25, 2005.

[55] M. Humbert, E. Ayday, J.-P. Hubaux, and A. Telenti. On non-cooperative genomic privacy. In R. Böhme and T. Okamoto, editors, *Financial Cryptography and Data Security*, volume 8975 of *Lecture Notes in Computer Science*, pages 407–426. Springer, 2015.

[56] P. Ilia, I. Polakis, E. Athanasopoulos, F. Maggi, and S. Ioannidis. Face/off: Preventing privacy leakage from photos in social networks. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 781–792, 2015.

[57] Z. Jiang, C. Heng, and B. Choi. Research note - Privacy concerns and privacy-protective behavior in synchronous online social interactions. *Information Systems Research*, 24(3):579–595, 2013.

[58] R. Johnson. Trade-off analysis of consumer values. *Journal of Marketing Research*, pages 121–127, 1974.

[59] A. Joinson. Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology*, 31(2):177–192, 2001.

[60] D. Kahneman, J. Knetsch, and R. Thaler. Fairness and the assumptions of economics. *Journal of Business*, 59(4):S285–S300, 1986.

[61] P. Kelley, L. Cranor, and N. Sadeh. Privacy as part of the app decision-making process. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pages 3393–3402, 2013.

[62] H. Krasnova, T. Hildebrand, and O. Guenther. Investigating the value of privacy in online social networks: Conjoint analysis. In *Proceedings of the International Conference on Information Systems (ICIS)*, 2009.

[63] H. Krasnova and N. Veltri. Privacy calculus on social networking sites: Explorative evidence from Germany and USA. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS)*, 2010.

[64] A. Lampinen, V. Lehtinen, A. Lehmuskallio, and S. Tamminen. We're in it together: Interpersonal management of disclosure in social network services. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3217–3226, 2011.

[65] M. Metzger. Privacy, trust, and disclosure: Exploring barriers to electronic commerce. *Journal of Computer-Mediated Communication*, 9(4), 2004.

[66] G. Milne and M.-E. Boza. Trust and concern in consumers' perceptions of marketing information management practices. *Journal of Interactive Marketing*, 13(1):5–24, 1999.

[67] T. Morlok. Sharing is (not) caring - The role of external privacy in users' information disclosure behaviors on social network sites. In *Proceedings of the Pacific Asia Conference on Information Systems (PACIS)*, 2016.

[68] H. Nissenbaum. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, 2009.

[69] A. Nochenson, J. Grossklags, and K. Lambert. Conducting an internet measurement project in an interdiscipinary class context: A case study. In *Proceedings of the 6th International Conference of Education, Research and Innovation*, pages 6938–6947, 2013.

[70] G. Nowak and J. Phelps. Understanding privacy concerns: An assessment of consumers' information-related knowledge and beliefs. *Journal of Direct Marketing*, 6(4):28–39, 1992.

[71] J. Nunnally. *Psychometric theory*. McGraw-Hill, 1967.

[72] A.-M. Olteanu, K. Huguenin, R. Shokri, M. Humbert, and J.-P. Hubaux. Quantifying interdependent privacy risks with location data. *Rapport LAAS n16018*, 2016.

[73] Y. Park, S. Campbell, and N. Kwak. Affect, cognition and reward: Predictors of privacy protection online. *Computers in Human Behavior*, 28(3):1019–1027, 2012.

[74] P. Pavlou. Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. *International Journal of Electronic Commerce*, 7(3):101–134, 2003.

[75] P. Pavlou and D. Gefen. Psychological contract violation in online marketplaces: Antecedents, consequences, and moderating role. *Information Systems Research*, 16(4):372–399, 2005.

[76] G. Petracca, A.-A. Reineh, Y. Sun, J. Grossklags, and

T. Jaeger. Aware: Preventing abuse of privacy-sensitive sensors via operation bindings. In *Proceedings of the 26th USENIX Security Symposium*, 2017.

[77] J. Phelps, G. Nowak, and E. Ferrell. Privacy concerns and consumer willingness to provide personal information. *Journal of Public Policy & Marketing*, 19(1):27–41, 2000.

[78] V. Prasnikar and A. Roth. Considerations of fairness and strategy: Experimental data from sequential games. *The Quarterly Journal of Economics*, 10(3):865–888, 1992.

[79] S. Preibusch. The value of privacy in web search. In *Proceedings of the Twelfth Workshop on the Economics of Information Security (WEIS)*, 2013.

[80] Y. Pu and J. Grossklags. An economic model and simulation results of app adoption decisions on networks with interdependent privacy consequences. In R. Poovendran and W. Saad, editors, *Decision and Game Theory for Security*, pages 246–265. Springer, 2014.

[81] Y. Pu and J. Grossklags. Using conjoint analysis to investigate the value of interdependent privacy in social app adoption scenarios. In *Proceedings of the International Conference on Information Systems (ICIS)*, 2015.

[82] Y. Pu and J. Grossklags. Sharing is caring, or callous? In *International Conference on Cryptology and Network Security*, pages 670–680. Springer, 2016.

[83] Y. Pu and J. Grossklags. Towards a model on the factors influencing social app users' valuation of interdependent privacy. *Proceedings on Privacy Enhancing Technologies*, 2016(2):61–81, 2016.

[84] H. Qian and C. Scott. Anonymity and self-disclosure on weblogs. *Journal of Computer-Mediated Communication*, 12(4):1428–1451, 2007.

[85] R. Radner and A. Schotter. The sealed-bid mechanism: An experimental study. *Journal of Economic Theory*, 48(1):179–220, 1989.

[86] P. Randall. Actively caring about the actively caring survey: Evaluating the reliability and validity of a measure of dispositional altruism. *Electronic Theses and Dissertations*, 2013.

[87] P. Rossi. bayesm: Bayesian inference for marketing/micro-econometrics. *URL http://CRAN. R-project. org/package= bayesm. R package version*, 2015.

[88] P. Rossi and G. Allenby. Bayesian statistics and marketing. *Marketing Science*, 22(3):304–328, 2003.

[89] A. Roth. Bargaining experiments. In J. Kagel, A. Roth, and J. Hey, editors, *The Handbook of Experimental Economics*, pages 253–348. Princeton University Press, 1995.

[90] A. Roth and M. Malouf. Scale changes and shared information in bargaining: An experimental study. *Mathematical Social Sciences*, 3(2):157–177, 1982.

[91] A. Savikhin and R. Sheremeta. Simultaneous decision-making in competitive and cooperative environments. *Economic Inquiry*, 51(2):1311–1323, 2013.

[92] C. Scott. Benefits and drawbacks of anonymous online communication: Legal challenges and communicative recommendations. *Free Speech Yearbook*, 41(1):127–141, 2004.

[93] S. Siegel and L. Fouraker. *Bargaining and group decision making: Experiments in bilateral monopoly*. McGraw-Hill, 1960.

[94] J. Smith, T. Dinev, and H. Xu. Information privacy research: An interdisciplinary review. *MIS Quarterly*, 35(4):989–1016, 2011.

[95] J. Smith, S. Milberg, and S. Burke. Information privacy: Measuring individuals' concerns about organizational practices. *MIS Quarterly*, 20(2):167–196, 1996.

[96] A. Soetevent. Anonymity in giving in a natural context - A field experiment in 30 churches. *Journal of Public Economics*, 89(11–12):2301–2323, 2005.

[97] S. Spiekermann, J. Grossklags, and B. Berendt. E-privacy in 2nd generation e-commerce: Privacy preferences versus actual behavior. In *Proceedings of the 3rd ACM Conference on Electronic Commerce*, pages 38–47, 2001.

[98] D. Stahl and E. Haruvy. Other-regarding preferences: Egalitarian warm glow, empathy, and group size. *Journal of Economic Behavior & Organization*, 61(1):20–41, 2006.

[99] J. E. Stiglitz. *Economics of the public sector*. W.W. Norton & Company, 2000.

[100] E. Stone and D. Stone. Privacy in organizations: Theoretical issues, research findings, and protection mechanisms. *Research in Personnel and Human Resources Management*, 8(3):349–411, 1990.

[101] J. Such and N. Criado. Resolving multi-party privacy conflicts in social media. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1851–1863, 2016.

[102] J. Such and M. Rovatsos. Privacy policy negotiation in social media. *ACM Transactions on Autonomous and Adaptive Systems*, 11(1):4:1–4:29, 2016.

[103] K. Thomas, C. Grier, and D. Nicol. unfriendly: Multi-party privacy risks in social networks. In *International Symposium on Privacy Enhancing Technologies*, pages 236–252. Springer, 2010.

[104] K. Train. *Discrete choice methods with simulation*, volume 8. Cambridge University Press, 2002.

[105] J. Tsai, S. Egelman, L. Cranor, and A. Acquisti. The effect of online privacy information on purchasing behavior: An experimental study. *Information Systems Research*, 22(2):254–268, 2011.

[106] J. Turow, C. Hoofnagle, D. Mulligan, N. Good, and J. Grossklags. The Federal Trade Commission and consumer privacy in the coming decade. *I/S: A Journal of Law and Policy for the Information Society*, 3(3):723–749, 2007.

[107] N. Wang, J. Grossklags, and H. Xu. An online experiment of privacy authorization dialogues for social applications. In *Proceedings of the 16th ACM Conference on Computer Supported Cooperative Work (CSCW)*, pages 261–272, 2013.

[108] N. Wang, P. Wisniewski, H. Xu, and J. Grossklags. Designing the default privacy settings for Facebook applications. In *Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 249–252, 2014.

[109] J. Weidman, W. Aurite, and J. Grossklags. Understanding interdependent privacy concerns and likely use factors for genetic testing: A vignette study. In *Proceedings of the 3rd International Workshop Genome Privacy and Security (GenoPri)*, 2016.

[110] R. Wheeler. Package algdesign: Algorithmic experimental design, 2010.

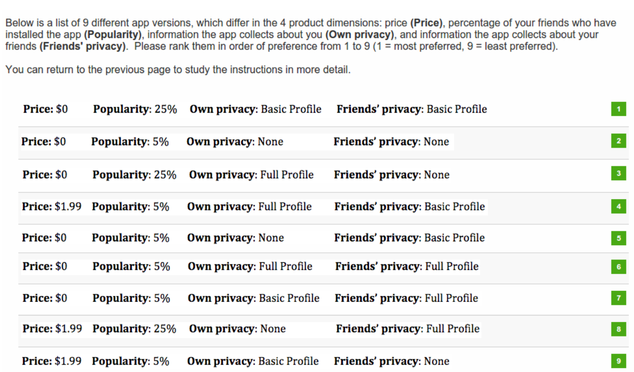
[111] P. Wisniewski, N. Islam, H. Richter Lipford, and D. Wilson. Framing and measuring multi-dimensional interpersonal

privacy preferences of social networking site users. *Communications of the Association for Information Systems*, 38(1):235–258, 2016.
[112] P. Wisniewski, H. Lipford, and D. Wilson. Fighting for my space: Coping mechanisms for SNS boundary regulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 609–618, 2012.
[113] D. Wittink and P. Cattin. Commercial use of conjoint analysis: An update. *The Journal of Marketing*, 53(3):91–96, 1989.
[114] H. Xu. The effects of self-construal and perceived control on privacy concerns. *Proceedings of the International Conference on Information Systems (ICIS)*, 2007.
[115] H. Xu, T. Dinev, J. Smith, and P. Hart. Information privacy concerns: Linking individual perceptions with institutional privacy assurances. *Journal of the Association for Information Systems*, 12(12):798, 2011.
[116] H. Xu, N. Wang, and J. Grossklags. Privacy by redesign: Alleviating privacy concerns for third-party apps. In *Proceedings of the 33rd International Conference on Information Systems*, 2012.

APPENDIX

A. A TYPICAL TASK IN FULL-PROFILE CONJOINT ANALYSIS STUDY

Figure 6 shows a task participants are expected to complete, if we implement our study by utilizing a full-profile conjoint analysis method [83]. Specifically, in this task, participants are required to rank 9 versions of an app according to their own preferences. Given that these app versions differ in 4 attributes, which further have multiple levels, this full-profile conjoint analysis task represents a higher cognitive challenge compared to the task we used in the current study (see Figure 1).


Below is a list of 9 different app versions, which differ in the 4 product dimensions: price **(Price)**, percentage of your friends who have installed the app **(Popularity)**, information the app collects about you **(Own privacy)**, and information the app collects about your friends **(Friends' privacy)**. Please rank them in order of preference from 1 to 9 (1 = most preferred, 9 = least preferred).

You can return to the previous page to study the instructions in more detail.

Price: $0	**Popularity:** 25%	**Own privacy:** Basic Profile	**Friends' privacy:** Basic Profile	1
Price: $0	**Popularity:** 5%	**Own privacy:** None	**Friends' privacy:** None	2
Price: $0	**Popularity:** 25%	**Own privacy:** Full Profile	**Friends' privacy:** None	3
Price: $1.99	**Popularity:** 5%	**Own privacy:** Full Profile	**Friends' privacy:** Basic Profile	4
Price: $0	**Popularity:** 5%	**Own privacy:** None	**Friends' privacy:** Basic Profile	5
Price: $0	**Popularity:** 5%	**Own privacy:** Full Profile	**Friends' privacy:** Full Profile	6
Price: $0	**Popularity:** 5%	**Own privacy:** Basic Profile	**Friends' privacy:** Full Profile	7
Price: $1.99	**Popularity:** 25%	**Own privacy:** None	**Friends' privacy:** Full Profile	8
Price: $1.99	**Popularity:** 5%	**Own privacy:** Basic Profile	**Friends' privacy:** None	9

Figure 6: A typical task in full-profile conjoint analysis study

B. AN EXAMPLE OF CALCULATING MONETARY VALUE OF PRIVACY

Following the practice in Krasnova et al. [62], we calculate privacy valuations based on utility associated with each attribute level. For example, consider Table 4 which lists the utilities a person has for each attribute level. We can then calculate the monetary value that person assigns to friends' basic information by taking the following four steps:

1. Calculating utility change of price level change from "$1.99" to "$0.00", which is: $1.63-(-1.63)=3.26$.

Attributes	Attribute Levels	Part-worth Utilities
Price	$0.00	1.63
	$1.99	-1.63
Network Popularity	5%	-0.73
	25%	0.73
Own Privacy	None	0.93
	Basic profile	0.43
	Full profile	-1.36
Friends' Privacy	None	0.60
	Basic profile	0.40
	Full profile	-1.00

Table 4: An example of part-worth utilities

2. Calculating amount of utility change per dollar change, which is $3.26/1.99=1.64$.
3. Calculating utility change of friends' privacy level change from "Basic profile" to "None", which is $0.6-0.4=0.2$.
4. Calculating dollar equivalent for friends' privacy level change from "Basic profile" to "None", i.e., dollar value of friends' basic information, which is $0.2/1.64=0.12$.

C. SURVEY INSTRUMENTS

Table 5, on the following page, includes the survey instruments that we utilized in our study.

Construct	Question wording
TrustInApp	Third-party app developers tell the truth about the collection and use of personal information.
	Third-party app developers can be relied on to keep their promises.
	I trust that third-party app developers will not use users' information for any irrelevant purposes.
	I can count on third-party app developers to take security measures to protect customers' personal information from unauthorized disclosure or misuse.
Privacy Knowledge	Companies today have the ability to place online advertisements that target you based on information collected about your web browsing behavior.
	When you go to a website, it can collect information about you even if you do not register.
	Popular search engine sites, such as Google, track the sites you come from and go to.
	Many of the most popular third-party apps reveal users' information to other parties, such as advertising and Internet tracking companies.
PastPrivacy Invasion	How often have you personally been victim online of what you felt was an invasion of privacy?
	How often have you personally been victim online of what you felt was an invasion of privacy?
	How often have you noticed others being victims online of what you felt was an invasion of privacy?
	How often have you noticed others being victims offline of what you felt was an invasion of privacy?
DispositionTo ValuePrivacy	Compared to others, I am more sensitive about the way personal information is handled.
	Keeping information private is the most important thing to me.
	Compared to others, I tend to be more concerned about threats to information privacy.
OtherRegarding Preference	I have recently helped a person with a problem.
	I should go out of my way to help people more often.
	If a member of my "social group" comes to me with a personal problem, I'm willing to listen without being judgmental.
	If a member of my "social group" needs help on a task, I am willing to help even if it causes me some inconvenience.
	I am willing to help a "social group" member I don't know.
Perceived Control	I believe I have control over who can get access to my personal information collected by third-party app developers.
	I think I have control over what my personal information is released by third-party app developers.
	I believe I have control over how my personal information is used by third-party app developers.
	I believe I can control my personal information provided to third-party app developers.
OwnPrivacy Concern	It usually bothers me when third-party app developers ask me for personal information.
	When third-party app developers ask me for personal information, I sometimes think twice before providing it.
	It bothers me to give my personal information to so many third-party app developers.
	I'm concerned that third-party app developers are collecting too much personal information about me.
FriendPrivacy Concern	It usually bothers me when third-party app developers ask me for my friends' personal information.
	When third-party app developers ask me for my friends' personal information, I sometimes think twice before providing it.
	It bothers me to give my friends' personal information to so many third-party app developers.
	I'm concerned that third-party app developers are collecting too much personal information about my friends.

Table 5: Survey instruments

Self-Driving Cars and Data Collection: Privacy Perceptions of Networked Autonomous Vehicles

Cara Bloom Joshua Tan Javed Ramjohn Lujo Bauer
Carnegie Mellon University
{cbloom, jstan, lbauer, jramjohn}@andrew.cmu.edu

ABSTRACT

Self-driving vehicles and other networked autonomous robots use sophisticated sensors to capture continuous data about the surrounding environment. In the public spaces where autonomous vehicles operate there is little reasonable expectation of privacy and no notice or choice given, raising privacy questions. To improve the acceptance of networked autonomous vehicles and to facilitate the development of technological and policy mechanisms to protect privacy, public expectations and concerns must first be investigated. In a study (n=302) of residents in cities with and without Uber autonomous vehicle fleets, we explore people's conceptions of the sensing and analysis capabilities of self-driving vehicles; their comfort with the different capabilities; and the effort, if any, to which they would be willing to go to opt out of data collection. We find that 54% of participants would spend more than five minutes using an online system to opt out of identifiable data collection. In addition, secondary use scenarios such as recognition, identification, and tracking of individuals and their vehicles were associated with low likelihood ratings and high discomfort. Surprisingly, those who thought secondary use scenarios were more likely were more comfortable with those scenarios. We discuss the implications of our results for understanding the unique challenges of this new technology and recommend industry guidelines to protect privacy.

1. INTRODUCTION

Networked autonomous robots in the form of drone swarms and commercial autonomous vehicles (AVs) are being researched, tested, and deployed. This technology is set to fundamentally shift common daily practices such as the use and ownership of automobiles [52]. At the time of data collection, Uber's self-driving car fleet had been deployed in Pittsburgh, PA for five months and was planning to expand to other states. The fleet is large enough that seeing the AVs has become quotidian to residents.

Two ethical concerns with the growing prevalence of AVs have received significant attention in the media and academic discourse. Ethical decision making—especially concerns with life-or-death decisions made by AVs—has been a major focus and has influenced public willingness to accept AVs as decision makers [41]. Commercial drivers, labor economists, and corporations have focused on the market effects of robots taking human jobs, both positive and negative [53]. A third ethical concern, and the focus of this paper, is the privacy-invasive capabilities of AVs, as well as the potential security risks associated with AV data collection. This ethical concern has received very little attention relative to decision-making and labor market changes.

Commercial fleets of networked AVs have the capability to collect location and movement data about residents of an entire city simply by storing the information already captured by their many sensors and using available software to analyze it. This capability poses a new regulatory conundrum, as it combines four different aspects of privacy-invasive technologies: (1) the ubiquitous capture of data in public, (2) physical surveillance by a privately owned company, (3) the ability to scale without additional infrastructure, and (4) the difficulty of notice and choice about data practices for physical sensors that capture data about non-users. Ubiquitous data collection in public has been implemented by cities such as London [48], which has sparked public debate over the efficacy and morality of surveillance. While cities are beholden to their constituents and residents, companies are beholden to their shareholders [13]. If a city like London and a company like Uber have the same data set of geo-temporal points, the former has an obligation to use it to better its constituents and the latter has an obligation to monetize it, bettering its shareholders.

While similar issues also apply to CCTV and dashboard cameras, the scalability and potential ubiquity of a networked self-driving car fleet is remarkable. Unlike CCTV, AVs can increase the bounds of their surveillance without additional infrastructure and can cover any public roads they are legally permitted to drive on. They use public infrastructure and are not reliant on privately owned property. The networked aspect differentiates them from dashboard cameras or individual self-driving cars (such as future Fords or Teslas) due to the scale of data collection and analytic capabilities on such aggregated sensor data.

These vehicles operate in public spaces where individuals do not have a reasonable expectation of privacy, and where notice and choice would be difficult to provide. Internet of things (IoT) devices such as Alexa already suffer from the difficulty of sufficiently notifying users of data collection (no

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

screen means no terms and conditions to read on the device), a task made more difficult when devices collect data from non-users [40]. Autonomous vehicles on public roads are constantly interacting with non-users: people who have not consented to any data collection or use and, as demonstrated in our study, have not yet thought about the potential privacy impacts.

As with many powerful new technologies, the large-scale capture and analysis of data enabled by AVs could lead to both benefits to the public and concerns. Ubiquitous sensing capabilities could be used to find Amber or Silver Alert citizens [48], but the same technology could also be used for less altruistic purposes. Insurers might analyze license plate logs to find out whether a customer speeds on the highway and adjust her car insurance rates accordingly, countries could identify dissidents, or an employee of the AV company could use the technology to stalk a celebrity or an ex-girlfriend, as Uber employees were found doing [7]. AV sensors could log the physical movements of every person within the purview of the fleet, making it possible to find anyone anywhere. The chilling effects of such surveillance and related dangers of ubiquitous data collection are well documented in privacy and security literature [39].

Reasonable expectation, unfairness, and deception are central themes for privacy regulation in the United States, so one key question is: where does the public draw the line between acceptable and unacceptable practices for autonomous networked robots? New technologies such as the Internet and more recently the Internet of Things can outpace the creation of reasonable privacy standards as they are quickly integrated into people's daily lives, leading to many inventive studies of the gaps in protection and how to patch them (e.g., [28, 17]). Users of these technologies can become habituated to the lack of privacy [51], making usable, effective privacy protections more difficult to enact. Therefore, it is important to explore privacy conceptions and strategies for privacy protection during the earliest phases of a new technology's implementation, before deployment outpaces the incorporation of privacy.

Whereas other potentially privacy-invasive technologies have required users opt in, AVs cannot give all pedestrians and drivers they encounter notice and choice. Companies operating such fleets could potentially offer notice outside of the information capture environment, but it would be difficult to give people the choice to opt out of information collection in all forms. Some information would have to be collected during the opt-out process, such as a license plate number to opt out of license plate recognition. Other options such as an opt-in process, privacy policies that limit the use of collected data, or even the removal of identifiable markers from stored data, are possible approaches. To make recommendations to the few companies currently operating in the space of networked AVs, privacy conceptions about the technology and its potential uses must first be understood.

Our investigation aims to fill this gap by exploring conceptions of the sensing and analysis capabilities of AVs; people's comfort with the different capabilities; and the effort, if any, to which they would be willing to go to opt out of data collection. We ran an online study of 302 participants using scenarios of increasing privacy invasiveness to measure how likely participants thought different potential capabilities of self-driving vehicles are, and how comfortable they are with those capabilities. Scenarios were framed using the Uber self-driving car fleet as an example. We recruited in Pittsburgh where the fleet has been deployed since September 2016 in addition to four other cities to investigate whether exposure to the technology changed conceptions or sentiments.

In addition to questions about likelihood and comfort with privacy capabilities, participants answered questions about general AV technology concerns like safety, their exposure to self-driving cars, bias against Uber, and demographic information. Responses were analyzed to determine likelihood and comfort levels as well as the relationship between likelihood, comfort, and potential explanatory variables.

We found that participants consider primary uses of AV sensors such as data collection, aggregation, storage, and analysis by the cars to be likely, and that participants express moderate comfort with these scenarios. Secondary use scenarios such as the recognition, identification, and tracking of individuals or their vehicles received the lowest ratings of likelihood and highest discomfort. Surprisingly, participants who thought the technology was more likely to have a privacy-invasive capability such as tracking were more likely to be comfortable with that capability. Though participants rated many capabilities likely and expressed high levels of discomfort, only one out of three would spend more than 10 minutes using an online opt-out system.

Pittsburgh participants who had exposure to the Uber self-driving car fleet (over 60% had seen one compared to 3% for other cities) were not statistically different in their conceptions of likelihood and comfort from residents of other cities who had never seen a self-driving car. The only factor that showed a significant increase in opt-out time was whether participants had received the privacy scenario priming questions, which participants noted had raised difficult questions they had not considered before. If public attention surrounding AVs expands from safety and employment issues to privacy issues, our findings suggest that peoples' overall comfort with AVs may increase, but so might privacy-seeking behavior as well. Understanding the complex privacy concerns in this space is essential for developing industry practices and regulation.

2. BACKGROUND AND RELATED WORK

The classic work in the area of AVs and privacy discusses the privacy implications for owners and users of AVs in detail and alludes to surveillance, noting that "[networked] autonomous vehicles could enable mass surveillance in the form of comprehensive, detailed tracking of all autonomous vehicles and their users at all times and places." The work focuses solely on the passengers within an AV who have ostensibly agreed to the terms and conditions, legally relinquishing their privacy the same way consumers do when using Google Maps [16]. In this paper we assess the more complex privacy concerns of those who interact with AVs, but are not necessarily users of the system. We next review consumer perception of AVs, followed by their technological capabilities and relevant regulations.

2.1 Consumer Perception

Research into consumer perceptions of AVs has examined general interest, trust in the cars' reliability and safety, and consumer feelings about how self-driving cars could impact the job market. Our work is one of the few that focuses on consumer privacy concerns and preferences regarding self-driving cars.

Consumer perception has been a popular area of discussion and research, given its potential impact on sales and market adoption. With AVs being deployed in test locations and viable plans to bring them to mass-market, studies have been conducted to gauge consumer interest. Schoettle and Sivak found that people are generally uninformed and had both high expectations about the benefits of autonomous vehicles and high levels of concern about riding in them. Additional concerns were changes to the job market, security against hacking, and data privacy concerns such as location and destination tracking [41]. This was one of the only studies of AVs that discussed data privacy, and it was not one of the central research questions.

In a Kelley Blue Book study (n>2000, weighted to census figures), 62% of participants did not want a world with solely AVs, with resistance decreasing with age [26], a trend corroborated by other studies of autonomy and age [1]. While these results shed insight into consumer preferences, this study was potentially biased by the extremity of its scenario, presenting a world with only autonomous cars to participants who likely live in an environment with only human-driven cars. This resistance to self-driving cars has been reinforced by other studies without extremity bias [10, 35].

Not all research studies have corroborated these findings. A survey (n=1517) run by AlixPartners found that three-quarters of U.S. drivers would want a self-driving car during their daily commute [36], a much higher level of acceptance than other studies had found. AlixPartners claims that prior surveys injected bias by placing emphasis on worst-case scenarios and that theirs found a balance that mitigated this bias.

Existing studies focus on consumer perception within the context of AVs, rather than the general public who are impacted just by being in the vicinity of AVs. Very little work seeks to study public perception decoupled from the framing of eventual consumption of self-driving cars. One such study is the MIT Moral Machine. It presents scenarios that show moral dilemmas "where a driverless car must choose the lesser of two evils, such as killing two passengers or five pedestrians" [31]. That study concerns the potential impact on and comfort of those in close proximity to an AV, but focuses solely on ethical issues related to physical safety.

Another study by Sleeper et al. explored perceptions of and comfort with vehicle-based sensing and recording used for purposes such as automatic lane correction and adaptive braking and cruise control. That study used hypothetical scenarios to examine perceived comfort for people who indirectly interact with vehicle sensors, include bystanders and nearby drivers. The authors found that perceived comfort with vehicle sensors increased when the benefits of the vehicle sensing was clear, particularly when benefits were related to safety [42]. In contrast to that study, our study explores perceptions and comfort with networked autonomous vehicles capable of large-scale data collection and analysis.

The body of research exploring consumer perceptions of AVs does have a consensus in one area: there is reluctance among the public toward accepting self-driving cars and issues of trust need to be addressed [47]. The focus is on potential consumers, rather than the public; safety concerns, rather than privacy concerns; and on individual AVs rather than commercial fleets of networked vehicles. Our study hopes to fill these gaps in understanding, especially because deployment of a commercial fleet has preceded private ownership of fully autonomous vehicles.

2.2 Technological Capability

Autonomous vehicles require extensive data in order to operate effectively. Their sensors typically include: GPS for navigation; a wheel encoder for monitoring the movements of the car; radar on the front and rear bumpers for identifying traffic; a camera near the rear-view mirror for color identification; lane departure, read collision, and pedestrian alerts; and a spinning light detection and ranging (LiDAR) sensor on the roof used for generating a 3D map of the environment (Figure 1) [20, 38].

The cameras bring up the greatest privacy concerns, especially if captured information is aggregated and centrally stored. A conceptual analogy used by our pilot study participants is CCTV surveillance. Thirteen states forbid the use of CCTV surveillance and all states require proper notice [12]. There are two flaws in this comparison: (1) CCTV is intended for surveillance while the sensors on a car are intended for autonomous driving, and (2) unlike CCTV, which is confined to a set space, AVs could be on any public road at any time. A more apt analogy could be the dashboard camera or 'dash cam,' yet information collected by dash cams is unlikely to be stored and analyzed centrally. It is unclear whether comfort with either CCTV or dash cams would translate to comfort with information capture by commercial fleets of AVs.

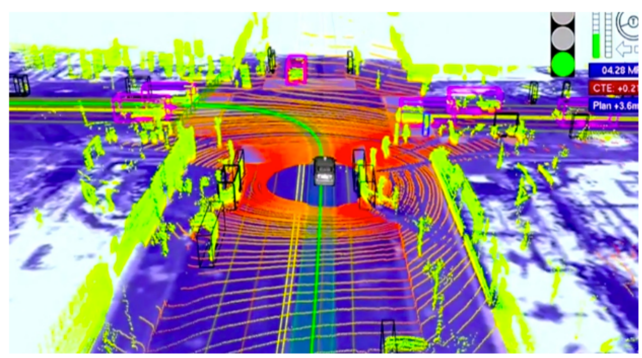


Figure 1: LiDAR Point Detection Heatmap

We suspect spinning LiDAR is the most foreign piece of sensing technology on an AV, for most people. LiDAR can be used for detecting and tracking objects; however, it is currently unable to identify individual people [38]. LiDAR data can potentially be combined with other data sources for concerning uses such as identifying how many people are at a protest.

Technological solutions aimed at mitigating AV privacy risks are not common, perhaps due to the lack of data surrounding consumer privacy preferences. Martinez-Balleste et al. describe ways to incorporate privacy-enhancing technology into the "smart city" by introducing the notion of "citizen privacy," explained as the right of privacy for entire communities [30]. The researchers provide models aimed toward technologies that collect vast amounts of data in large, public settings; these models can be used to inform analysis of the privacy implications for AVs which can be thought of as an element of a smart city.

Self-driving cars can be seen as a new privacy invasive technology capable of always-on monitoring during operation, yet no notice is currently provided about how captured data will be used. Through our work, we gauge what the public deems acceptable, in an effort to inform the industry about what practices their potential customers and government stakeholders could want.

2.3 Regulation

Despite the tendency of U.S. law to react slowly to technological advances, the federal government has been convening stakeholders and developing regulatory principles for AVs. While there has been much discussion, there has been little movement on formal federal regulations and legislation. The Government Accountability Office (GAO) analyzed ten auto companies with regard to their location-data services and found that each had moved towards recommended privacy practices [49]. A year after the GAO report, the Alliance of Automobile Manufacturers submitted a statement to the Federal Trade Commission (FTC) with a commitment from the member companies to uphold privacy principles, specifically the traditional Fair Information Privacy Practices (FIPPs) [3]. In 2016, a National Highway Traffic Safety Administration report reiterated existing privacy stances by the government, such as notice and choice, desire to encourage competition in the realm of privacy, and the need to secure data [50].

Recently, state policymakers have taken steps to address AV privacy concerns [18]. The State of California passed a law that requires manufacturers of AVs to provide written disclosure of what data is collected [44], prompting backlash from the automotive industry [15]. As of 2016, 20 states have introduced autonomous car regulation, and since 2012, 34 states and D.C. have considered autonomous car legislation [33]. Eleven of these states have passed legislation, with two states using executive orders to mandate policy. While the California law is generally cited by the media, Michigan was highlighted as the first state to pass comprehensive AV regulations [4]. The legislation focused less on privacy constraints and instead legalized self-driving ride-sharing services, allowing for truck platoons, autonomous cars without drivers, and testing and usage on public roads [4]. The only major restriction, which states like Georgia, Maryland, Illinois, and Tennessee, have also introduced, is that the deployment of autonomous vehicles on public property is limited to automakers, requiring companies like Uber, Lyft, and Google to work with automakers in order to deploy vehicles [5].

At the federal level there is no binding legislation that addresses the privacy concerns associated with AVs. The Center for Internet and Society at Stanford maintains a wiki with current legislative and regulatory actions in the space of cyber law [43]; as of March 2017, the only enacted legislation with reference to AVs is the Fixing America Surface Transportation (FAST) Act. This legislation only instructs the GAO to "assess the status of autonomous transportation technology policy developed by U.S. public entities" [24]. Interestingly, the only other federal bill listed was the Autonomous Vehicle Privacy Protection Act of 2015. Unfortunately, this bill is still in committee deliberations by the House Transportation and Infrastructure Subcommittee on Highways and Transits [25]. The bill is not yet fully developed, only stating that the GAO needs to provide a public report assessing the ability of the Department of Transportation to address autonomous vehicle issues like consumer privacy—almost the same provision as in the FAST Act [25].

The closest regulations to data collection by the many sensors and cameras on an AV are those for dash cams. Legal authors Stitilis and Laurinaitis recognize that privacy is a huge concern with dash cams and hold that the benefits do not necessarily outweigh the harms. They relate back to the traditional view of privacy as the right to be left alone and cite EU laws that guarantee the right to privacy. Even with simple dash cam footage, it is difficult to balance priorities—cams are difficult to ban and people in public spaces do not have a reasonable expectation of privacy [45]. Deleting dash cam footage can be considered evidence tampering, which raises the question of if self-driving cars record information, would retention be necessary for legal compliance [23]? Despite the lack of uniform regulations, dash cams appear to have more privacy regulations than AVs at the state level, where some states prohibit recording when the owner is not driving and prohibit using them to surreptitiously record audio while being hidden from plain sight [23]. Publication of the collected data involves separate regulation and public perception, though cases generally involve simple uses such as determining the cause of an accident [21].

Another precedent regarding pedestrian privacy is the controversy surrounding Google Street View. The Street View technology was met with substantial scrutiny, with accusations about failing to properly blur faces and collecting excessive data, such as Wi-Fi signals [46]. Despite its use in public spaces the use of automated technology to collect data about people and their behaviors prompted considerable anxiety and response from the company [11].

Greenblatt asserts that the law has not prepared for the emergence of self-driving cars and will not be ready [19]. Given the deployment of AVs and lack of federal legislation, along with a mixed response from the states, Greenblatt appears to be right. Much like the rest of the Internet of Things, technology has outpaced the law, which, especially in the realm of privacy and security, has led to deficiencies that have damaged consumer trust in IoT devices [14]. If AVs follow the same direction as IoT devices have, the trust the public has in self-driving cars could be damaged by a major privacy or security breach—hampering the adoption of the technology and potentially inviting unwanted regulation. Our work hopes to provide the industry with guidance for crafting privacy protections that build trust rather than break it.

Journalists have investigated the extent of data collection and tracking features in high-tech cars, with mixed results. Articles have speculated car companies collect more than they say [6]. Companies are quick to respond, but often do not assuage privacy concerns or disclose data collection pro-

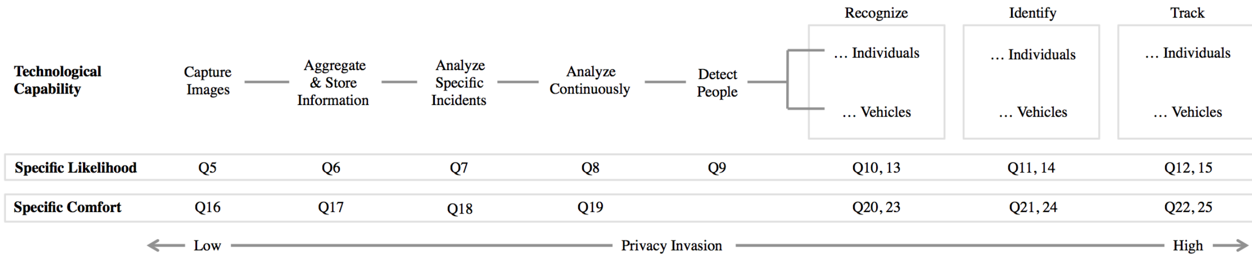


Figure 2: Privacy Scenarios

cedures [32, 22]. Policies enacted by ride-sharing companies have become the standard for the industry, a self-regulatory approach that currently defines most of U.S. privacy law [22]. Automotive data collection, in AVs or otherwise, presents a new set of privacy challenges for the industry. Companies such as BMW "[have] been inundated with requests from advertisers and technology companies to get their hands on vehicle data" [37]. The potential uses of and inferences from vehicle data by advertisers are extensive, even more so for AVs which necessarily collect a greater amount of data.

There are few legal protections for pedestrians and drivers against the capture and use of images taken by AVs. Even when their images are taken without their knowledge and consent, the current legal protection of tort law is limited when the likenesses of pedestrians are captured in photographs in a public environment, such as a city street [23].

3. METHODOLOGY

Our work seeks to understand public perceptions surrounding autonomous vehicles from a privacy standpoint, focusing on potentially privacy invasive capabilities of commercial AV fleets. We designed an online survey to explore people's conceptions of the sensing and analysis capabilities of self-driving vehicles; their comfort with the different capabilities; and the effort, if any, to which they would be willing to go to opt out of data collection.

3.1 Survey Questionnaire

As the first study explicitly investigating privacy conceptions surrounding networked fleets of AVs, an exploratory survey was chosen as the research method with the goal of identifying what consumers think is reasonable for AVs to do.

Participants were asked to focus on a fleet of self-driving cars operated by a single company that shares information with each other as well as the company, rather than single individually owned cars which have different capabilities and associated concerns. Only sensors on the outside of the car were to be considered, not any within the vehicle or any corresponding mobile application, to limit unknown effects. As a quality check, participants chose either "Yes, I understand" or "No, I did not read the short text" to move on to the next section.

The survey structure split participants into two groups to control for the priming effect of privacy questions. The Primed group received the set of scenario questions represented visually in Figure 2 to investigate conceptions of the sensing and analysis capabilities of self-driving vehicles (Specific Likelihood questions, Q16-25) as well as a set to gauge comfort level with the technological capabilities (Specific Comfort questions, Q5-15). The Unprimed group skipped these two sections and began with a set of General Comfort questions (Q26-35), which are represented visually in Figure 3. Two scenarios in this set concerned privacy. Eight other scenarios were included to both obfuscate the privacy questions and to facilitate comparison of discomfort due to privacy reasons with discomfort due to other aspects of the technology (e.g. safety or job market concerns). Both the Primed group and Unprimed group answered the General Comfort questions, the latter responding after answering the two specific question sets.

All scenario questions were piloted iteratively and discussed with pilot participants, who fell into one of four groups: non-technical, university students, security and/or privacy students, and robotics students. Pilots with the latter two groups developed the content and validity of scenarios to accurately fit the technology and accomplish research goals. Additional pilots were done to increase understanding of the scenarios. A small-scale pilot (n=41) using online recruiting was run to gather preliminary data, then final minor edits were made using data from these responses.

Specific Likelihood Questions

Participants were asked to answer questions about their conceptions of the current technological capabilities of AV fleets. These questions were designed to identify what people thought AV fleets were already doing. Participants rated different scenarios on a five point Likert scale from "Strongly Disagree" to "Strongly Agree." The scenarios began with those we assessed as least privacy invasive (i.e. image capture) and increased in invasiveness to scenarios involving recognition, identification, and tracking of people and vehicles. To help participants understand complex privacy concepts, examples were provided using the context of the Uber self-driving car fleet. Scenarios and examples can be found in Appendix A Q5-15 and are demonstrated visually in Figure 2.

Specific Comfort Questions

After the likelihood questions, participants in the Primed group indicated their comfort level with the same technological capabilities on a five point Likert scale from "Very Uncomfortable" to "Very Comfortable." While the Specific Likelihood questions measured what participants thought was realistically occurring, the Specific Comfort questions

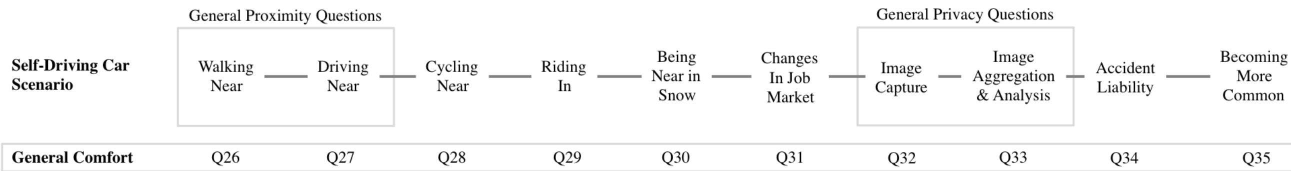


Figure 3: General Scenarios

asked participants how comfortable they would be if the technological capability was realized. By using the same capability scenarios, the relationship between conceptions of likelihood and comfort could be explored, as well as the difference in comfort at different levels of technological capability. Explanatory examples for these questions encouraged participants to imagine the technological capability in an every-day scenario and can be found in Appendix A Q16-25.

General Comfort with Self-Driving Cars

The General Comfort questions (Q26-34) covered general concerns with AVs identified by Schoettle and Sivak [41], including walking and driving near AVs, changes in the job market, and legal liability in an accident (Figure 3). Within the General Comfort set were two privacy questions (Q32-33), one concerning image capture and the other concerning aggregation and analysis of captured images. Using this set of scenarios, the effect of being primed with the likelihood and comfort scenarios could be measured for both general comfort with fleets of AVs and privacy concerns.

Quantification of Effort to Opt-out

The scenario-based questions measured conceptions and attitudes about AV technology, not behavior. To investigate whether discomfort would lead to action, we asked the following question:

> Q36. Suppose the company operating the fleet of self-driving cars has implemented a system so pedestrians and drivers can opt out of data collection by the cars. By going through their online system, people can have images of them blurred so their identity is protected and they cannot be tracked. How many minutes would you spend in the system to successfully opt out?

Response options were grouped into five-minute buckets for times between one and thirty minutes with options for "0 minutes" and "More than 30 minutes."

Exposure and Bias

We investigated the effects of using Uber as the example for our study, fearing that Uber as an example would bias results due to the many news stories circulating about the company during data collection and their strong effect on public opinion [7]. However, feedback from pilot participants indicated that even if Uber had not been used as the example, many participants would have still used the Uber self-driving car fleet as a mental model. Using Uber consistently kept some participants from using it as a mental model when others did not, which made the biasing effects easier to detect and measure. Additionally, using AVs that were already deployed in public spaces and familiar to many of our participants' made it more likely that they would be able to accurately envision and have developed opinions about the scenarios that we cover in our survey.

To measure the bias created by the use of Uber as an example, we asked participants to express their agreement with five statements on a five point Likert scale from "Strongly Disagree" to "Strongly Agree." The questions (Q36-40) assessed topics such as whether they would have answered the questions differently if Uber had not been used as an example. In addition to the bias questions, exposure to the technology and interaction with Uber were measured (Q45). Exposure questions included whether participants had read an article about Uber self-driving cars or ridden in one; interaction questions included whether participants used the Uber app or had protested against Uber.

Participant Characteristics

To further understand participants and the role characteristics play in their conceptions of networked AVs, demographic information was collected including gender, age, educational experience, and industry. Technical experience and general privacy attitudes were also recorded, the latter using the IUIPC question framework [29]. Email addresses were only collected to distribute compensation.

3.2 Recruitment

Participants were recruited from five cities of similar size and demographics: Pittsburgh, PA; Cleveland, OH; Cincinnati, OH; Rochester, NY; and Buffalo, NY. Participants were recruited in all five cities using local Craigslist ads and posts on city-specific Reddit forums. Posters were also used to recruit in six major central neighborhoods of Pittsburgh. Multiple methods were used to avoid bias from any one type of respondent and participants outside the specified cities were disqualified. Tracking of recruitment method was done via unique survey links. Participants who finished the survey could choose to give their email address to be entered into a random drawing for one of six $50 Amazon Gift Cards. The survey was run for two weeks beginning February 16 and closing March 3, 2017.

3.3 Analysis

We performed hypothesis tests to understand the relationship between participants' perceptions of likelihood and comfort with AV technological capabilities. We test the correlation between participants' perceived likelihood and comfort with specific self-driving car capabilities using Spearman's ρ. To understand whether perceived likelihood ratings differed between person- and vehicle-specific capabilities, as well as how these ratings differed between different groups of participants, we binned likelihood ratings into {likely, very likely} and {very unlikely, unlikely, neither likely nor unlikely} and use Fisher's exact test. Comfort ratings were similarly tested using {uncomfortable, very uncomfortable} and {very uncomfortable, uncomfortable, neither uncomfortable nor comfortable} bins. In addition, we tested whether participants' specified opt-out minutes differed between participant segments using the Mann-Whitney U test.

All hypothesis tests used a significance level of $\alpha = 0.05$. For general self-driving comfort ratings, opt-out minutes, and comfort with specific AV capabilities, we performed exploratory testing with respect to many variables. To account for this, we applied the Holm-Bonferonni method within each family of tests and report corrected p-values.

4. RESULTS

Of the 312 survey responses, 248 gave complete responses and ten were excluded. Participants were excluded for failing the attention-check question (two participants), entering a location outside the scope of the study (one), or because they were Uber employees (seven). These last were excluded due to concerns about the lack of generalizability from their data to other populations. Additionally, multiple Uber employees seemed to be taking the survey only to see what the questions were, as they chose the neutral option for every Likert question and did not enter an email address for the gift card raffle.

Our sample was slightly skewed by the recruitment methods. Over half of participants (55%) were recruited via Reddit, which led to the sample being more male, technically experienced, and younger than the general population, due to the demographics of Reddit users [9]. Of the participants who answered demographic questions, 61% identified as male. The average age was 34 years, ranging from 18 to 79, and 24% were majoring in or had a degree or job in computer science, computer engineering, information technology, or a related field. The sample was more well-educated than the population with 13 with professional or doctoral degrees (5%), 45 with masters degrees (18%), 108 with bachelors degrees (43%), 16 with an associates degree, 49 with some college experience (19%), and 21 participants who had no college experience (8%). Based on the IUIPC privacy questions, the overwhelming majority of participants had strong beliefs concerning their own privacy. It should be noted though, that these questions were given at the end of the survey which had already raised many privacy concerns and could have increased participants' privacy sentiments.

Participants were randomly assigned to either the Primed or Unprimed group. The Primed group had 158 (52%) participants and the Unprimed group had 144 (48%). Of the five recruitment locations, the largest sample came from Pittsburgh (200, 68%), followed by Cleveland (63, 21%).

4.1 Exposure and Bias

Participants indicated their experience with Uber's AV technology in the survey by checking any of the fourteen statements that applied to them, seen in Figure 4. Statements

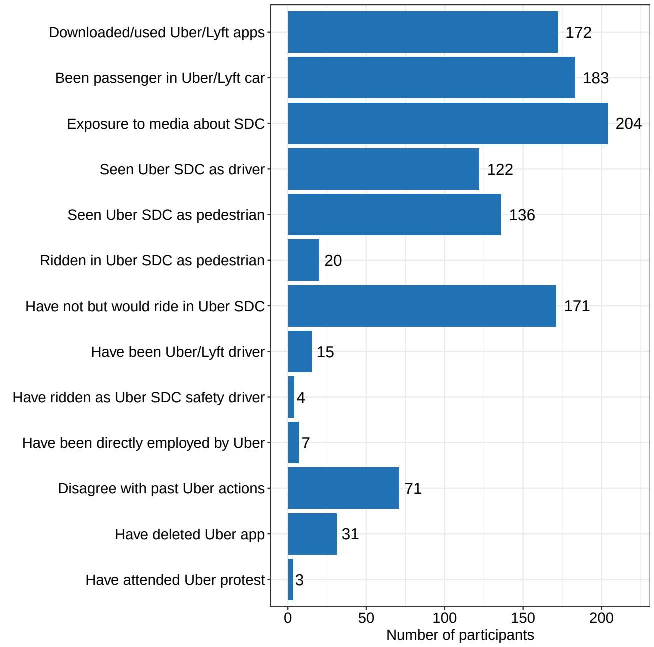


Figure 4: Participant exposure to Uber and self-driving cars (SDC).

covered not only exposure to self-driving cars, but also exposure to ride sharing technology, attitudes towards self-driving cars, attitudes towards Uber, behaviors indicative of negative opinions of Uber, and employment status by Uber or another ride-sharing service. Participants who self-identified as Uber employees (excluding drivers) are included in Figure 4 but were excluded from all other analyses; there were seven Uber employees in the study including four who had ridden as "safety drivers" in Uber self-driving cars.

Participants recruited from Pittsburgh had higher rates of exposure to ride sharing and self-driving technology in all areas. Notably, 78% of Pittsburgh participants and 42% of non-Pittsburgh participants had read an article, viewed a program, or learned online about Uber self-driving vehicles, indicating a high level of exposure to information about Uber's AV technology prior to this study. Seven percent of Pittsburgh participants had already ridden in a self-driving car and 64% had seen one as a pedestrian, compared to <1% and 3% of non-Pittsburgh participants, respectively. These results suggest that Pittsburgh residents generally have high exposure to the technology itself while residents of cities without the Uber self-driving car fleet have little to no exposure, though there may be some response bias where Pittsburghers with greater exposure were more likely to take our survey.

Negative attitudes towards Uber and associated behaviors were also prevalent in our sample, which we tested because of the many public controversies associated with the company. Twenty-three percent of all participants disagreed with actions Uber had taken and 10% had deleted the Uber mobile app. Three participants (<1%) had participated in protests against Uber. Most importantly, due to biases or preconceived notions, 14% of participants agreed and 3% strongly agreed that if Uber had not been used as an example they would have answered the survey questions differently. Furthermore, 18% would trust another self-driving car company over Uber to have their best interests in mind, indicating that some of the distrust is company-related, not directly related to AV technology.

Since Uber is the most visible company currently operating networked fleets of autonomous robots in public spaces and despite its controversies, it was logical to use Uber as an example in scenarios. We decided that the ecological validity and use of a single mental model outweighed incurred bias. As one participant added in the free-text response, "I would have automatically used Uber in my own mind as an example." If some participants had used Uber as their mental model, while others used Google or Tesla, interpretation of our results would be more difficult. Consistent use of Uber as an example standardized the context for all participants and allowed us to ask participants about a technology that was already deployed in their city or cities like theirs.

4.2 Conceptions of Technological Capabilities

The trend in ratings of likelihood was inversely related to how privacy invasive the Specific Likelihood Question was, as ranked by researchers and shown in Figure 5.

Participants overwhelmingly rated basic capabilities such as image capture and aggregated storage as likely to be occurring, 87% and 91% respectively. Detection of humans was rated as likely by a similar proportion, at 88%. Under the assumption that images were already captured and stored, 94% of participants thought analysis for specific incidents, such as traffic accidents, was likely and 88% thought it was likely information was analyzed continuously for general tasks such as navigation. These are primary uses that directly impact the function of AVs. We found a clear division in ratings of capability between primary and secondary uses, where secondary uses are uses not necessary for the primary function of the AV. The secondary uses we explored are identification, recognition, and tracking of individuals and vehicles. Participants found primary uses to be highly likely, yet no more than half of participants rated each secondary use scenario as likely. Likelihood ratings for secondary uses are summarized in Table 1. Notably, the scenario that received the lowest likelihood rating by participants was also one of the most privacy invasive as ranked by coauthors: identification of individuals at 22%.

Overall there was a clear delineation in ratings of likelihood between primary and secondary use scenarios. Due to lack of information about the capabilities Uber self-driving cars actually have, only two scenarios are known to be occurring: image capture and detection of people. Almost 9 out of 10 participants accurately thought these verifiable scenarios were likely, as expected. A substantial minority of participants, no fewer than 1 out of 5, believed that even the most privacy invasive scenarios were likely to be occurring. While most participants held that primary uses were likely and secondary uses were not, many thought that the AV technology was being used to the extent of its capability in extremely privacy invasive ways, such as identifying pedestrians.

4.3 Comfort and Privacy Preferences

Discomfort level with each of the Specific Comfort questions (Q16-25) was quantified using the proportion of participants who chose "Uncomfortable" or "Very Uncomfortable." Participants were generally more comfortable with primary uses than with secondary uses. Discomfort was lowest for the least privacy invasive scenario (image capture, 16%) and highest for one of the most privacy invasive scenarios (tracking of vehicles, 85%). Generally high levels of discomfort were seen with: image storage (42%), analysis of specific incidents (36%), and continuous analysis (43%). The example used for the incident analysis scenario was Uber reviewing images captured of an accident, which could have explained why the associated discomfort was lower; as P95 noted in her free response, "If I have an accident with a driverless car, the recording is something useful, but that in my opinion should be the only reason the recordings/information should be released." Participants could have viewed this scenario as similar to dash cameras, which have known benefits and accepted norms of behavior. Of the secondary use scenarios, more than half of participants were uncomfortable with every scenario except vehicle recognition (43%), which was notably also the scenario rated most likely.

Scenario	Individuals	Vehicles
Recognition	38% (53)	46% (64)
Identification	22% (31)	28% (38)
Tracking	42% (58)	34% (47)

Table 1: Perceived likelihood of secondary use scenarios. The percentage (count) of participants that saw a scenario as likely or very likely are shown.

Comfort levels tended to decrease as questions increased in privacy invasiveness. The proportion of participants uncomfortable with aggregated storage was statistically significantly greater than with just image collection (Fisher's Exact Test, 42% vs. 15%, $p<0.001$). For secondary use scenarios—recognition, identification, and tracking—participants were more comfortable with recognition than identification or tracking. In particular, participants expressed higher discomfort with tracking of vehicles than identification of vehicles (85% vs. 71%, $p=0.040$) and higher discomfort with identification than recognition for both vehicles and individuals (71% vs. 43% for vehicles, 76% vs. 54% for individuals, $p<0.002$ for both). Notably, we did not observe statistically significant differences in comfort between continuous analysis and analysis of specific events (43% vs. 36%) nor between identification and tracking of individuals (both 76%). We also did not observe statistically significant differences in comfort for the three secondary use scenarios between individuals and vehicles.

4.3.1 Relationship Between Likelihood and Comfort

We also investigated whether rating a given capability scenario as likely was correlated with comfort with that same scenario. We found that there was a statistically significant positive correlation between likelihood and comfort ratings for identification (Spearman's $\rho=0.28$, $p=0.001$) and tracking ($\rho=0.17$, $p=0.049$) of individuals; and recognition ($\rho=0.19$, $p=0.028$), identification ($\rho=0.30$, $p<0.001$), and tracking ($\rho=0.22$, $p=0.019$) of cars. Likelihood and comfort ratings correlated most strongly for secondary use scenarios involving identification.

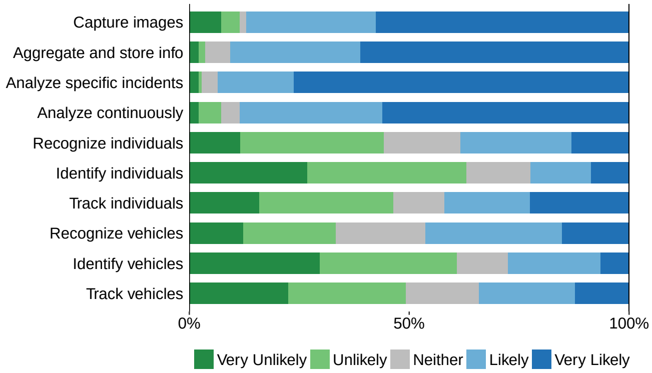


Figure 5: Likelihood ratings.

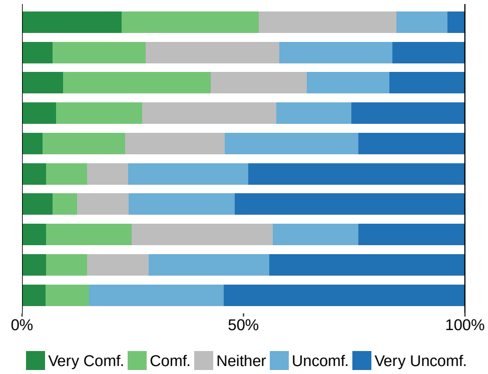


Figure 6: Comfort ratings.

Scenario	Overall	PGH	Non-PGH
Capture images	16% (20)	14% (13)	19% (7)
Aggregate and store info	42% (54)	43% (40)	38% (14)
Analyze specific incidents	36% (46)	36% (33)	35% (13)
Analyze continuously	43% (55)	39% (36)	51% (19)
Recognize individuals	54% (70)	57% (52)	49% (18)
Identify individuals	76% (98)	75% (69)	78% (29)
Track individuals	76% (98)	78% (72)	70% (26)
Recognize vehicles	43% (56)	46% (42)	38% (14)
Identify vehicles	71% (92)	68% (63)	78% (29)
Track vehicles	85% (95)	84% (67)	88% (28)

Table 2: Discomfort with technological capabilities in different scenarios, overall and by whether participants lived in Pittsburgh. The percentage (count) of participants that were uncomfortable or very uncomfortable with a scenario are shown.

Based on Cohen's guidelines we can interpret the strength of these observed relationships between likelihood and comfort ratings [8]. We observed moderate and near-moderate positive correlations between likelihood and comfort for vehicle identification and individual identification, respectively. Among participants that thought vehicle identification was likely, 59% were uncomfortable with this capability. In contrast, among those that thought vehicle identification was not likely, 76% were uncomfortable.

For the remaining scenarios, we observed small to moderate effect sizes (from $\rho = 0.17$ for individual tracking, to $\rho = 0.22$ for vehicle tracking). In all cases the direction of the correlation is positive; increased likelihood ratings were associated with increased comfort ratings.

We were surprised to find positive correlations between likelihood conceptions and comfort. Participants were more comfortable with a capability if they thought it was likely happening. It was expected that participants who thought a particular capability was already occurring would be more uncomfortable with it because they would feel more pressing concern with a technology that is already in use. Instead the opposite was observed, with higher ratings of likelihood related to higher levels of comfort.

4.3.2 *Other Factors Related to Comfort*

A number of additional factors were tested for effect on discomfort. Rather than use the Specific Comfort questions presented only to the Primed group, this analysis used the General Comfort questions shown to all participants as the measure of comfort with self-driving cars. In this exploratory analysis the intention was to uncover variables that could explain what made participants more or less comfortable with basic privacy invasive capabilities of AVs, namely image capture and analysis (Figure 3, Q32-33).

To obfuscate and compare these privacy questions, other concerns with AVs were also measured. Four scenarios in the General Comfort questions (Figure 3, Q26-35) concerned discomfort with proximity to a self-driving car: walking near (24% expressed discomfort), driving near (25%), cycling near (49%), and being near one in the snow (61%). Other causes of discomfort identified by Schoettle and Sivak's large-sample survey [41] were also explored. Forty-four percent of participants were uncomfortable with changes in the job market due to self-driving cars and 72% were uncomfortable with legal liability resulting from an accident with a self-driving car. We also asked one general question about how comfortable participants felt about self-driving cars becoming more common, to which 30% indicated discomfort.

The privacy scenarios (Q32-33) in the General Comfort questions made more participants feel uncomfortable than any other scenario: 85% were uncomfortable with image capture of people and license plates, and 77% were uncomfortable with that data being aggregated and analyzed. It was surprising that more participants felt uncomfortable with image capture than data aggregation and analysis, but a possible explanation is that participants are more uncomfortable with the fact that data is collected than potential uses of those data. Participants in pilot studies had difficulty articulating negative outcomes from image analysis, which could also explain the observations.

Using the Mann-Whitney U test, the following variables were investigated for a difference in discomfort: priming with privacy scenarios, Pittsburgh residence, gender, technical experience, and bias against Uber. As an example, we tested whether comfort with image capture statistically significantly differed between Pittsburgh and non-Pittsburgh participants. The Kruskal-Wallis test was used to compare discomfort between overall conceptions of likelihood and age range. Overall conceptions of likelihood were quantified as a numerical score (0-11) representing the number of scenarios from the Specific Likelihood questions a participant had found likely. Surprisingly, none of the explanatory variables we explored had a statistically significant impact on discomfort.

To determine if these variables did not explain differences in privacy comfort, or if they did not explain any difference in comfort, the same tests were run on comfort with proximity to self-driving cars and comfort with these cars becoming more common in general. Discomfort with proximity as a driver and proximity as a pedestrian were quantified for each participant as whether they were uncomfortable with both, one, or neither of the scenarios (0-2).

We found statistically significant differences in discomfort with proximity to self-driving cars between participants with different discomfort levels for the Specific Comfort scenarios (Kruskal-Wallis, $\chi^2(10) = 42.28$, $p<0.001$). Similarly, we found statistically significant differences in discomfort with self-driving car technology becoming more common between discomfort levels with specific scenarios ($\chi^2(10) = 35.32$, $p<0.001$). In both cases, however, we did not observe a clear trend relating discomfort (with proximity or with the technology becoming more common) and overall discomfort with specific scenarios.

Whether participants had technical experience explained statistically significant differences in comfort with self-driving car technology becoming more common in general (Mann-Whitney $U = 4506.5$, $p=0.049$). Technical experience—studying or employed in computer science, computer engineering, information technology, or related—was related to increased comfort with the technology becoming more common (technical: 17% uncomfortable, non-technical: 34% uncomfortable), but it did not explain comfort with either privacy-related scenario (image capture or analysis).

The survey did not ask participants why they were uncomfortable with any specific scenario, so it is possible that the reason participants expressed discomfort with proximity is in fact because of privacy invasion and not for safety reasons. In this case having higher concern with proximity could be explained by discomfort with the sensors, not the possibility of being endangered, which is not corroborated by the dominance of safety in public discourse surrounding the technology. It is also possible that the lack of statistical significance for the two privacy questions within the General Comfort questions set could be due to a high baseline discomfort level.

4.3.3 *Indications of Opt-Out Behavior*

The set of explanatory variables described in the previous section were investigated for their effect on how long participants were willing to spend in an online system in order to opt out of identifiable data collection. Nine percent of participants would not use the online system, 37% would spend 5 minutes or fewer, 22% would spend 6-10 minutes, 20% would spend 11-30 minutes, and 12% would spend more than half an hour. Priming with the specific scenario questions was the only variable for which we observed statistically significantly different opt-out times (Mann-Whitney $U = 9847.5$, $p=0.022$), with opt out times higher for the Primed group (primed median: 6-10 minutes, non-primed median: 1-5 minutes).

This difference can be partially explained by the open text responses participants chose to give at the end of the survey. Four thoughtful responses discussed the opt out question specifically, three of whom disagreed with the idea of opting out, arguing instead that people should opt in or simply not have identifiable information captured. These responses showed nuanced thought about the nature of the technology and privacy implications which another participant (P91) noted had "raised issues [she] had never even considered." The nature of the scenario questions given to the Primed group presented scenarios and privacy implications that pilot study participants said they had not thought of before the study. Simply posing questions about potential privacy invasive scenarios increased the amount of time participants would spend to mitigate such invasions. It also shows that when the public is made aware of potential privacy invasions without accurate information about actual data collection and use practices, there is an increase in privacy-seeking behavior.

5. DISCUSSION

This study explored a previously unknown space: technological and privacy perceptions surrounding networked AVs, specifically the Uber commercial fleet of self-driving vehicles. We identified what technological capabilities the public ascribed to fleets of self-driving cars, how comfortable they were with those capabilities, and the effort to which they would go to protect themselves from privacy invasion. What we found was a complex space where perceived likelihood correlated with higher comfort, attributes that we thought would predict attitude and behavior had no observed effect, and simply asking questions about potential privacy scenarios increased participants' predictions of the time they would spend to opt out. Nevertheless, findings gleaned from this study can be used to recommend industry strategy and practices to assuage discomfort, protect privacy, and increase acceptance of this new technology.

5.1 Limitations

Sampling and recruitment bias could have played a role in our results. Participants came only from mid-sized cities in the Midwestern and Mid-Eastern regions of the United States, which limits the generalizability to more urban or rural populations as well as other nations. This limitation was the result of a conscious design choice: we specifically wanted to focus on people who had experience with fleets of AVs, which meant recruiting in Pittsburgh; then to compare opinions of people who were significantly less exposed to self-driving vehicles, we chose cities geographically near and demographically similar to Pittsburgh so as to avoid additional confounds. Future studies should diversify to more urban and rural areas, as well as to other cultures. Comparisons between exposed and unexposed populations should be available soon, as Uber deploys fleets in cities like San Francisco and more rural areas such as Michigan [4].

Another limitation of this study is the format used to conduct it. An online survey allowed us to reach over 300 people and learn about their conceptions of AV technology, but it was limited in depth. Many variables that could explain comfort and inform policy are as of yet unidentified and unexplored. More in-depth research could also assess what costs and benefits people think can come from the surveillance capabilities of networked fleets of autonomous robots.

5.2 Privacy Conceptions

Using the scenarios concerning technological capability (Figure 2) we learned what the public thinks self-driving cars currently do and how they feel about it. As expected, participants overwhelmingly (and correctly) believed that AVs have the capability to gather rich information about their environment and detect humans, as well as that AV fleets can perform off-line analyses of the collected information. The majority of participants generally thought secondary uses of collected information such as identification and tracking were not likely, though these scenarios still had a substantial minority of participants (22% to 46%) rating them as likely. As expected, comfort significantly decreased as scenarios became more invasive and a division was found between primary and secondary uses. Secondary uses were differentiated by participants in pilots and free-text responses by their degree of necessity and invasion: the invasion was often found to be needlessly 'too far', whereas primary uses could be rationalized.

Surprisingly, for the secondary use scenarios, rather than higher conceptions of likelihood correlating with higher discomfort, we observed the opposite. Participants who rated a potentially privacy-invasive scenario as likely were more likely to be comfortable with that scenario; this might be explained by learned helplessness or resignment to perceived inevitability. Learned helplessness is when in negative situations where an individual has no ability to change the circumstances, such as the invasion of privacy by autonomous vehicles, people increasingly accept the situation as a coping mechanism. With no power to change the environmental factors that cause a negative response, the negative response itself is changed [54].

Similarly, if participants had perceived the technological capability as not only likely, but as normal or inevitable, this could have led to increased comfort. These findings support the need for research and privacy enhancing technologies and policies early in the technology's life cycle. As people become resigned over time, the deployment of AV technology may outpace restrictions, as previously mentioned in reference to IoT technology, making it harder to integrate privacy protections.

Causes of Discomfort

Though a participant's perceived likelihood of a particular scenario explained her comfort with that scenario, other expected explanatory variables did not. None of the explanatory variables tested explained any difference in comfort with AV image capture and analysis (Q32-33). In contrast, greater technical experience was associated with increased comfort with self-driving cars becoming more common in general. We expected that technical experience would have one of two potential effects: greater knowledge leading to a better understanding of potential negative impacts and consequences and hence more concern; or, alternatively, better understanding of the benefits and hence less concern. Support for the latter was found, but only for comfort with AV technology in general, not for comfort with privacy scenarios, where technical experience had no observable effect. A possible explanation is that comfort with AV technology in general is derived mainly from safety and employment concerns, rather than privacy concerns.

We expect that proximity concern is a combination of privacy concerns and safety concerns, with significantly greater weight given to safety than privacy based on the narrative of public discourse, open-text responses of participants, and the phrasing of the questions. In this case, privacy discomfort could be indicative of safety discomfort for other reasons, such as that they are both caused by an innate distrust of the technology. More nuanced exploration would be needed to answer these questions, perhaps via interview studies.

Time to Opt Out

Though high levels of discomfort with the different technological capabilities were found, half of participants would spend only five or fewer minutes using an online system to opt out of identifiable data collection by commercial autonomous vehicles. The only factor that explained a difference in opt-out time was whether the participant had been primed with specific privacy scenarios. Presenting people with scenarios that suggested the possibility of privacy invasion made people predict that they would spend more effort mitigating the privacy invasion. No other variables, including exposure to self-driving cars or bias against Uber, explained a difference in time to opt out.

Should the public be exposed to questions regarding privacy invasive capabilities, there could be an increased move towards privacy-seeking behavior such as opting out or perhaps protesting. Research and media attention is currently focused on safety and employment, but more of our participants were uncomfortable with privacy invasive capabilities than with either of these popular concerns. Even participants in the Unprimed group, who did not see questions regarding recognition, identification, and tracking, were more likely to be uncomfortable with privacy scenarios than with proximity scenarios. If public attention were to shift towards the third ethical concern—privacy—findings in this study indicate that discussions would reveal great discomfort and the act of discussing such concerns could cause a change in behavior concerning commercial self-driving vehicles.

5.3 Recommendations for Industry Practice

One of the central questions investigated by this study was where the public draws the line on acceptable and unacceptable privacy practices by companies operating networked autonomous vehicles in public spaces. The sentiments of participants tended toward acceptance of technologies they thought were being implemented as necessary components, but toward discomfort with secondary analysis of information such as recognition, identification, and tracking of people or vehicles. Additionally, participants would overwhelmingly use a system to opt out of identifiable information capture, though some expressed that an opt-out tool is unsuited to the technology.

The synthesis of these findings shows that people, regardless of their exposure to AV technologies, are uncomfortable with privacy-invasive secondary uses and, to a lesser extent, with primary uses such as continuous analysis of data captured by networked AVs. The only secondary use that could potentially be considered useful and acceptable was recognition of vehicles, which participants rationalized could be useful for taking extra precautions against erratic drivers. With other new technologies, the argument can be made that if the privacy intrusions conflict with individuals' preferences, they need not use that technology; but with sophisticated sensors operating in public places people have no practical ability to avoid information capture. It is then necessary that companies operating such fleets of AVs and other robots like drones either implement industry self-regulation or be regulated to protect the public. Our findings suggest that such regulation should focus on secondary data uses, with which the public is overwhelmingly uncomfortable and would actively avoid if given the opportunity.

Currently this regulation could take three forms: industry self-regulation, federal regulation, or state and local restrictions. The Alliance of Automotive Manufacturers has jurisdiction due to the necessity of autonomous vehicle companies partnering with traditional automotive companies [5] and this organization is committed to the Fair Information Privacy Practices (FIPPs) [2]. All of the foundational necessities are in place, but this organization has not yet applied them directly to AVs, or in particular to concerns raised by their external sensors. Federal regulation could take multiple forms; traditionally roads and cars are under the jurisdiction of the National Highway Transportation Safety Administration [34], though the FTC frequently crosses into other jurisdictions to enact privacy regulation. Both agencies support notice and choice, the first two FIPPs. Local and state governments are interfacing directly with these AV companies already though, and do require knowledge of their practices before allowing them access to public roads. These cities and states could set precedent for broader practice by working with the companies to create practices that balance the need for information with citizens' privacy. The companies themselves could create or adapt other privacy enhancing technologies such as face and license plate blurring, such as that done by Google Maps cars [16].

Additionally, it is in the best interest of companies operating AV fleets to be more transparent about their data collection and use practices. While the public has not yet considered the privacy implications of AV technology the way it has safety implications, this study found that bringing up privacy concerns causes people to be less comfortable with being near and utilizing self-driving car technology and to express intentions of actively mitigating privacy invasion. Such attitudes could cause increased backlash not only from the public, which has already been vocal about reservations about safety and employment, but from the city and state governments that are currently debating whether to allow autonomous vehicles to operate within their jurisdictions.

Safety concerns can be rebutted with the argument that the new technology (AVs) is less concerning than the current environment (human drivers), but companies like Uber cannot argue that data capture by networked autonomous vehicles is less concerning than the current environment where there are no networked vehicles capable of city-scale surveillance. Standard arguments for the technology are more difficult to apply and companies have yet to make a case for—or provide public services that—demonstrate data collection is net positive for the populations of the cities they operate in.

6. CONCLUSION

Our study investigated the largely unexplored space of privacy concerns surrounding autonomous vehicles. We found that participants generally thought networked fleets of autonomous vehicles were collecting and analyzing data about them, and that more than 40% thought this technology was already being used to track people's movements. Scenarios such as tracking and identification caused overwhelming discomfort, while participants expressed moderate discomfort with primary uses of data such as continuous analysis for navigation. If a participant thought a particular capability was likely to be occurring, she was more comfortable with that capability, perhaps because she thought it was normal or because she was resigned to it.

Surprisingly, privacy concerns caused higher proportions of participants to express discomfort than either of the more common concerns—physical proximity or changes in the job market. These feelings of discomfort with privacy-invasive capabilities were not explained by any of the variables we examined, indicating that attitudes were either too nuanced for detection by this study, were resistant to the effects of other variables, or were explained by unexplored additional factors. Interestingly, the amount of time participants predicted they would spend on privacy-protective behaviors was not as resistant: simply asking priming questions about autonomous vehicle capabilities increased participants' predictions of how long they would spend in an online system to opt out of identifiable data collection. Future studies can further investigate the relationship between priming, attitudes, and behaviors, and increase the understanding of privacy concern in this technological context.

Autonomous vehicle technology is set to become increasingly prevalent in the next decade and permanently alter daily life for millions of people [27]. Privacy research early in the development life cycle of this unique technology can be used to shape industry practices and regulation before intentional or unintentional privacy invasions become a part of the technology. It is important to investigate privacy implications of networked autonomous vehicles before deployment outpaces understanding of potential ramifications. We recommend policies differentiate between primary and secondary uses of sensor data, restricting secondary uses to preserve public privacy.

7. ACKNOWLEDGMENTS

This work was supported in part with a gift from Google. The authors would like to thank Matthew Bajzek and Aki Hitomi for help with early versions of this work.

8. REFERENCES

[1] H. Abraham, C. Lee, S. Brady, C. Fitzgerald, B. Mehler, B. Reimer, and J. F. Coughlin. Autonomous vehicles, trust, and driving alternatives: A survey of consumer preferences. Technical report, Massachusetts Institute of Technology, Cambridge, MA, 2016. `http://agelab.mit.edu/files/publications/2016_6_Autonomous_Vehicles_Consumer_Preferences.pdf`. Accessed March 2017.

[2] Alliance of Automobile Manufacturers. Automotive privacy, 2017. `https://autoalliance.org/connected-cars/automotive-privacy-2/`.

[3] Alliance of Automotive Manufacturers Inc. Letter to FTC, 2014. `https://autoalliance.org/connected-`

cars/automotive-privacy-2/letter-to-ftc/. Accessed October 2016.

[4] J. Bhuiyan. Michigan just became the first state to pass comprehensive self-driving regulations. *Recode*, Dec. 2016. https://www.recode.net/2016/12/9/13890080/michigan-dot-self-driving-cars-laws-automakers. Accessed March 2017.

[5] J. Bhuiyan. A series of U.S. state laws could prevent Uber or Google from operating self-driving cars. *Recode*, Feb. 2017. http://www.recode.net/2017/2/25/14738966/self-driving-laws-states-gm-car-makers. Accessed March 2017.

[6] J. M. Broder. That Tesla data: What it says and what it doesn't. *The New York Times*, 2013. http://wheels.blogs.nytimes.com/2013/02/14/that-tesla-data-what-it-says-and-what-it-doesnt. Accessed October 2016.

[7] B. Carson. Uber's unraveling: The stunning, 2 week string of blows that has upended the world's most valuable startup. *Business Insider*, Mar 2017. http://www.businessinsider.com/uber-scandal-recap-2017-3.

[8] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. L. Erlbaum Associates, 1988.

[9] M. Duggan and A. Smith. 6% of online adults are Reddit users. *Pew Internet & American Life Project*, 2013. http://www.pewinternet.org/2013/07/03/6-of-online-adults-are-reddit-users/.

[10] Ernst & Young. Autonomous vehicles: How much do we need?, 2016. http://www.ey.com/gl/en/industries/automotive/ey-autonomous-vehicles-how-much-human-do-we-need. Accessed March 2017.

[11] A. Eustace. WiFi data collection: An update, 2010. https://googleblog.blogspot.com/2010/05/wifi-data-collection-update.html.

[12] R. Fogel. CCTV and video surveillance laws in US, Dec. 2011. http://www.smartsign.com/blog/cctv-laws-in-us/. Accessed March 2017.

[13] R. E. Freeman. *Strategic management: A stakeholder approach*. Cambridge University Press, 2010.

[14] FTC Staff Report. Internet of Things privacy & security in a connected world. Technical report, The Federal Trade Commission, Jan. 2015. https://www.ftc.gov/system/files/documents/reports/federal-trade-commission-staff-report-november-2013-workshop-entitled-internet-things-privacy/150127iotrpt.pdf. Accessed March 2017.

[15] M. Geuss. Automakers balk at California's proposed self-driving car rules. *Ars Technica*, Oct. 2016. http://arstechnica.com/cars/2016/10/automakers-balk-at-californias-proposed-self-driving-car-rules/. Accessed October 2016.

[16] D. J. Glancy. Privacy in autonomous vehicles. *Santa Clara L. Rev.*, 52:1171, 2012.

[17] J. Gluck, F. Schaub, A. Friedman, H. Habib, N. Sadeh, L. F. Cranor, and Y. Agarwal. How short is too short? Implications of length and framing on the effectiveness of privacy notices. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 321–340, Denver, CO, 2016. USENIX Association.

[18] E. Goodman. Self-driving cars: Overlooking data privacy is a car crash waiting to happen. *The Guardian*, Aug. 2016. https://www.theguardian.com/technology/2016/jun/08/self-driving-car-legislation-drones-data-security. Accessed October 2016.

[19] N. Greenblatt. Self-driving cars will be ready before our laws are. *IEEE Spectrum*, Jan. 2016. http://spectrum.ieee.org/transportation/advanced-cars/selfdriving-cars-will-be-ready-before-our-laws-are. Accessed March 2017.

[20] E. Guizzo. How Google's self-driving car works. *IEEE Spectrum*, Oct. 2011. http://spectrum.ieee.org/automaton/robotics/artificial-intelligence/how-google-self-driving-car-works. Accessed March 2017.

[21] C. J. Junior, S. Muse, and C. Jung. Crowd analysis using computer vision techniques. *IEEE Signal Processing Magazine*, 27(5):66–67, 2010. http://ieeexplore.ieee.org/document/5562657/. Accessed October 2016.

[22] A. LaFrance. Driverless-car makers on privacy: Just trust us. *The Atlantic*, Mar. 2016. http://www.theatlantic.com/technology/archive/2016/03/self-driving-car-makers-on-privacy-just-trust-us/474903/. Accessed October 2016.

[23] S. Lehto. The surprising legal ramifications of having a dashcam in your car. *Road & Track*, Jan. 2017. http://www.roadandtrack.com/car-culture/a32124/the-surprising-legal-ramifications-of-having-a-dash-cam-in-your-car/. Accessed March 2017.

[24] Library of Congress. H.R 22 - FAST Act, Dec. 2015. https://www.congress.gov/bill/114th-congress/house-bill/22. Accessed March 2017.

[25] Library of Congress. H.R 3876 - Autonomous Vehicle Privacy Protection Act of 2015, Nov. 2015. https://www.congress.gov/bill/114th-congress/house-bill/3876. Accessed March 2017.

[26] T. Lien. Consumers aren't as excited as the auto industry about self-driving cars. *Los Angeles Times*, 2016. http://www.latimes.com/business/technology/la-fi-tn-kbb-self-driving-car-survey-20160927-snap-story.html. Accessed March 2017.

[27] T. Litman. Autonomous vehicle implementation predictions. *Victoria Transport Policy Institute*, 28, 2014. http://www.vtpi.org/avip.pdf.

[28] B. Liu, M. S. Andersen, F. Schaub, H. Almuhimedi, S. A. Zhang, N. Sadeh, Y. Agarwal, and A. 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, 2016. USENIX Association.

[29] N. K. Malhotra, S. S. Kim, and J. Agarwal. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information systems research*, 15(4):336–355, 2004.

[30] A. Martinez-Balleste, P. Perez-martinez, and A. Solanas. The pursuit of citizens' privacy: A privacy-aware smart city is possible. *IEEE Communications Magazine*, 51(6):136–141, 2013. http://ieeexplore.ieee.org/document/6525606/. Accessed October 2016.

[31] Massachusetts Institute of Technology. MIT Moral Machine, July 2016. http://moralmachine.mit.edu. Accessed March 2017.

[32] E. Musk. A most peculiar test drive, 2013. https://www.tesla.com/blog/most-peculiar-test-drive. Accessed October 2016.

[33] National Conference of State Legislature. Autonomous vehicles - self-driving vehicles enacted legislation, Feb. 2017. http://www.ncsl.org/research/transportation/autonomous-vehicles-self-driving-vehicles-enacted-legislation.aspx. Accessed March 2017.

[34] National Highway Traffic Safety Administration. Laws administered by NHTSA, 2017. https://www.nhtsa.gov/laws-regulations/statutory-authorities.

[35] K. Naughton. Billions are being invested in a robot that Americans don't want. *Bloomberg*, May 2016. https://www.bloomberg.com/news/articles/2016-05-04/billions-are-being-invested-in-a-robot-that-americans-don-t-want. Accessed March 2017.

[36] K. Naughton. Three-quarters of U.S. drivers say they'd cede wheel to robot. *Bloomberg*, June 2016. https://www.bloomberg.com/news/articles/2016-06-30/three-quarters-of-u-s-drivers-say-they-d-cede-wheel-to-robot. Accessed March 2017.

[37] C. Neiger. Advertisers are begging car companies for your data. *The Motley Fool*, Jan. 2015. http://www.fool.com/investing/general/2015/01/25/advertisers-are-begging-car-companies-for-your-dat.aspx. Accessed October 2016.

[38] J. Petit. Self-driving and connected cars: Fooling sensors and tracking drivers, 2015. https://www.blackhat.com/docs/eu-15/materials/eu-15-Petit-Self-Driving-And-Connected-Cars-Fooling-Sensors-And-Tracking-Drivers.pdf. Accessed March 2017.

[39] N. M. Richards. The dangers of surveillance. *Harvard Law Review*, 126(7):1934–1965, 2013.

[40] F. Schaub, R. Balebako, A. L. Durity, and L. F. Cranor. A design space for effective privacy notices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 1–17, Ottawa, 2015. USENIX Association.

[41] B. Schoettle and M. Sivak. A survey of public opinion about autonomous and self-driving vehicles in the US, the UK, and Australia. 2014.

[42] M. Sleeper, S. Schnorf, B. Kemler, and S. Consolvo. Attitudes toward vehicle-based sensing and recording. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 1017–1028, New York, NY, USA, 2015. ACM.

[43] Stanford Center for Internet and Society. Automated driving: Legislative and regulatory action. http://cyberlaw.stanford.edu/wiki/index.php/. Last Updated Feb. 2015. Accessed March 2017.

[44] State of California. Autonomous vehicles in California, 2016. https://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/bkgd. Accessed October 2016.

[45] D. Stitilis and M. Laurinaitis. Legal regulation of the use of dashboard cameras: Aspects of privacy protection. *Computer Law and Security Review*, 32(4):316–326, Apr. 2016. http://www.sciencedirect.com/science/article/pii/S0267364916300267. Accessed October 2016.

[46] D. Streitfeld. Google concedes that drive-by prying violated privacy. *The New York Times*, Mar. 2013. http://www.nytimes.com/2013/03/13/technology/google-pays-fine-over-street-view-privacy-breach.html. Accessed October 2016.

[47] C. Tennant, S. Howard, B. Franks, and M. Bauer. Autonomous vehicles: Negotiating a place on the road. Technical report, London School of Economics, 2016. http://www.lse.ac.uk/website-archive/newsAndMedia/PDF/AVs-negociating-a-place-on-the-road-1110.pdf. Accessed March 2017.

[48] M. J. Thomas. *Combining facial recognition, automatic license plate readers and closed-circuit television to create an interstate identification system for wanted subjects*. PhD thesis, Monterey, California: Naval Postgraduate School, 2015.

[49] United States Government Accountability Office. In-car location-based services, Dec. 2013. http://www.gao.gov/assets/660/659509.pdf. Accessed October 2016.

[50] U.S. Department of Transportation National Highway Traffic Safety Administration. Federated automated vehicles policy, 2016. http://www.nhtsa.gov/nhtsa/av/index.html. Accessed October 2016.

[51] Y. Wang, P. G. Leon, A. Acquisti, L. F. Cranor, A. Forget, and N. Sadeh. A field trial of privacy nudges for Facebook. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 2367–2376. ACM, 2014.

[52] Wilson and Mitchell. Beyond Uber, Volvo and Ford: Other automakers' plans for self-driving vehicles. *Los Angeles Times*, Aug 2016. http://www.latimes.com/business/autos/la-fi-automakers-self-driving-20160819-snap-htmlstory.html.

[53] M. Wisniewski. Driverless cars could improve safety, but impact on jobs, transit questioned. *Chicago Tribune*, Jul. 2016. http://www.chicagotribune.com/news/ct-driverless-cars-getting-around-20160703-story.html.

[54] C. B. Wortman and J. W. Brehm. Responses to uncontrollable outcomes: An integration of reactance theory and the learned helplessness model. *Advances in experimental social psychology*, 8:277–336, 1975.

APPENDIX

A. SURVEY

1. What city do you live in?

◯ Pittsburgh, PA
◯ Rochester, NY

◯ Buffalo, NY
◯ Cincinnati, OH
◯ Cleveland, OH
◯ Other - Write In (Required)

2. How did you learn about this survey?
◯ Poster
◯ Craigslist
◯ Reddit
◯ Word of mouth
◯ Other - Write In (Required)

Please read survey information carefully. This survey explores opinions about fleets of self-driving cars. It is NOT intended to test or judge your knowledge of self-driving car technology.

For this survey suppose: (1) A fleet of self-driving cars is operated in your city (2) The cars are owned and operated by a private company (3) The cars are networked to share information with each other and the company

One example of this is the Uber self-driving car fleet currently operated in Pittsburgh, PA.

This survey is NOT about: (1) Individually owned self-driving cars (2) Any sensors on the inside of the car

3. Do you understand what this survey is and is not about?
◯ No, I didn't read the short text (please read)
◯ Yes, I understand

4. Likelihood of Self-Driving Car Scenarios [Primed group only]

You will be presented with scenarios about a networked fleet of self-driving cars. Choose how likely you think the scenarios are to be happening now from 'very unlikely' to 'very likely.' Please read each question carefully.

5. Self-driving cars capture images of their surroundings
◯ Very Unlikely
◯ Unlikely
◯ Neither Unlikely nor Likely
◯ Likely
◯ Very Likely

6. Information captured by the self-driving cars is aggregated and stored
For example, Uber stores data collected by all of its self-driving cars in a central location
◯ Very Unlikely
◯ Unlikely
◯ Neither Unlikely nor Likely
◯ Likely
◯ Very Likely

7. Information captured by a self-driving car during a specific incident is analyzed by the operating company
For example: Images captured by an Uber self-driving car during a car accident are used by Uber to determine the cause
◯ Very Unlikely
◯ Unlikely
◯ Neither Unlikely nor Likely
◯ Likely
◯ Very Likely

8. Information captured continuously by the self-driving cars is analyzed
For example: Data collected by all Uber self-driving cars is used by Uber to understand weather conditions
◯ Very Unlikely
◯ Unlikely
◯ Neither Unlikely nor Likely
◯ Likely
◯ Very Likely

9. Self-driving cars detect humans
◯ Very Unlikely
◯ Unlikely
◯ Neither Unlikely nor Likely
◯ Likely
◯ Very Likely

10. A self-driving car recognizes a person that has been encountered before by a different self-driving car in the fleet
◯ Very Unlikely
◯ Unlikely
◯ Neither Unlikely nor Likely
◯ Likely
◯ Very Likely

11. Individuals are identified by name when they encounter one of the self-driving cars in the fleet
For example: Uber knows that the pedestrian next to one of its self-driving cars is Alice
◯ Very Unlikely
◯ Unlikely
◯ Neither Unlikely nor Likely
◯ Likely
◯ Very Likely

12. Individuals are tracked using each time they encounter one of its self-driving cars in the fleet
◯ Very Unlikely
◯ Unlikely
◯ Neither Unlikely nor Likely
◯ Likely
◯ Very Likely

13. A self-driving car recognizes a vehicle that has been seen by another self-driving car in the fleet
For example: Uber knows that different self-driving cars encountered the same vehicle on different days, but does not know who owns the vehicle
◯ Very Unlikely
◯ Unlikely
◯ Neither Unlikely nor Likely
◯ Likely
◯ Very Likely

14. Vehicle owners are identified by name when a vehicle encounters one of the self-driving cars in the fleet
For example: Uber knows that the minivan in front of one of its self-driving cars is owned by Alice ◯ Very Unlikely
◯ Unlikely
◯ Neither Unlikely nor Likely
◯ Likely

◯ Very Likely

15. Vehicles are tracked using each time they encounter one of the self-driving cars in the fleet
For example: Uber assembles a list with location, date, and time of each time self-driving cars encountered Alice's minivan
◯ Very Unlikely
◯ Unlikely
◯ Neither Unlikely nor Likely
◯ Likely
◯ Very Likely

Comfort with Self-Driving Cars [Primed group only]

Choose how comfortable you are with the scenarios from 'very uncomfortable' to 'very comfortable.' Please read each question carefully.

16. I would feel _____ if self-driving cars captured images of me (but did not store or analyze those images.)
For example: An Uber self-driving car captures an image of you in a crosswalk, then discards the image after it leaves the intersection.
◯ Very Uncomfortable
◯ Uncomfortable
◯ Neither Uncomfortable nor Comfortable
◯ Comfortable
◯ Very Comfortable

17. I would feel _____ if self-driving cars captured and stored images of me (but did not analyze those images)
For example: An Uber self-driving car captures an image of you in a crosswalk and it is stored on a computer with many similar images, but Uber does not use the images.
◯ Very Uncomfortable
◯ Uncomfortable
◯ Neither Uncomfortable nor Comfortable
◯ Comfortable
◯ Very Comfortable

18. I would feel _____ if self-driving cars captured images of me and analyzed images of specific events
For example: Uber analyzes specific images captured by a self-driving car (including images of you) to determine the cause of a traffic incident.
◯ Very Uncomfortable
◯ Uncomfortable
◯ Neither Uncomfortable nor Comfortable
◯ Comfortable
◯ Very Comfortable

19. I would feel _____ if self-driving cars captured images of me and analyzed images continuously
For example: Uber continuously analyzes images captured by all self-driving cars (including images of you) to gauge traffic conditions.
◯ Very Uncomfortable
◯ Uncomfortable
◯ Neither Uncomfortable nor Comfortable
◯ Comfortable
◯ Very Comfortable

20. I would feel _____ if each time I encountered a self-driving car, I was recognized from past encounters with other self-driving cars (but not by name).
For example: Uber knows that different self-driving cars encountered you in different locations on different days, but does not know who you are
◯ Very Uncomfortable
◯ Uncomfortable
◯ Neither Uncomfortable nor Comfortable
◯ Comfortable
◯ Very Comfortable

21. I would feel _____ if I was identified by images captured by a self-driving car
For example: An Uber self-driving car captures an image of your face as you cross the street and Uber links the image to your name
◯ Very Uncomfortable
◯ Uncomfortable
◯ Neither Uncomfortable nor Comfortable
◯ Comfortable
◯ Very Comfortable

22. I would feel _____ if I was tracked each time I encountered a self-driving car.
For example: Uber assembles a list with location, date, and time of each time you encounter a self-driving car.
◯ Very Uncomfortable
◯ Uncomfortable
◯ Neither Uncomfortable nor Comfortable
◯ Comfortable
◯ Very Comfortable

23. I would feel _____ if each time my car encountered a self-driving car, it was recognized from past encounters with other self-driving cars (but not by owner's name).
For example: Uber knows that different self-driving cars encountered your car in different locations on different days, but does not know who owns the car.
◯ Very Uncomfortable
◯ Uncomfortable
◯ Neither Uncomfortable nor Comfortable
◯ Comfortable
◯ Very Comfortable

24. I would feel _____ if my car was identified by images captured by a self-driving car
For example: An Uber self-driving car captures an image of your license plate as you drive and Uber uses the links the license plate to your name
◯ Very Uncomfortable
◯ Uncomfortable
◯ Neither Uncomfortable nor Comfortable
◯ Comfortable
◯ Very Comfortable

25. I would feel _____ if my car was tracked each time it encountered a self-driving car.
For example: Uber assembles a list with location, date, and time of each time your car encounters a self-driving car.
◯ Very Uncomfortable
◯ Uncomfortable
◯ Neither Uncomfortable nor Comfortable
◯ Comfortable
◯ Very Comfortable

General Self-Driving Car Questions

You will be presented with scenarios about a networked fleet of self-driving cars. Choose how comfortable you are with the scenarios from 'very unlikely' to 'very likely.' Please read each question carefully.

26. I would feel _____ walking near a self-driving car.
◯ Very Uncomfortable
◯ Uncomfortable
◯ Neither Uncomfortable nor Comfortable
◯ Comfortable
◯ Very Comfortable

27. I would feel _____ driving near a self-driving car.
◯ Very Uncomfortable
◯ Uncomfortable
◯ Neither Uncomfortable nor Comfortable
◯ Comfortable
◯ Very Comfortable

28. I would feel _____ cycling near a self-driving car.
◯ Very Uncomfortable
◯ Uncomfortable
◯ Neither Uncomfortable nor Comfortable
◯ Comfortable
◯ Very Comfortable

29. I would feel _____ riding in a self-driving car.
◯ Very Uncomfortable
◯ Uncomfortable
◯ Neither Uncomfortable nor Comfortable
◯ Comfortable
◯ Very Comfortable

30. I would feel _____ being near a self-driving car in the snow.
◯ Very Uncomfortable
◯ Uncomfortable
◯ Neither Uncomfortable nor Comfortable
◯ Comfortable
◯ Very Comfortable

31. I would feel _____ about the changes in the job market due to self-driving cars.
◯ Very Uncomfortable
◯ Uncomfortable
◯ Neither Uncomfortable nor Comfortable
◯ Comfortable
◯ Very Comfortable

32. I would feel _____ if a self-driving car captured pictures of me and my license plate.
◯ Very Uncomfortable
◯ Uncomfortable
◯ Neither Uncomfortable nor Comfortable
◯ Comfortable
◯ Very Comfortable

33. I would feel _____ if images captured by self-driving cars were aggregated and analyzed
◯ Very Uncomfortable
◯ Uncomfortable
◯ Neither Uncomfortable nor Comfortable
◯ Comfortable
◯ Very Comfortable

34. I would feel _____ about legal liability in an accident with a self-driving car.
◯ Very Uncomfortable
◯ Uncomfortable
◯ Neither Uncomfortable nor Comfortable
◯ Comfortable
◯ Very Comfortable

34. I would feel _____ about networked fleets of self-driving cars becoming more common in general.
◯ Very Uncomfortable
◯ Uncomfortable
◯ Neither Uncomfortable nor Comfortable
◯ Comfortable
◯ Very Comfortable

Opting Out of Information Capture

Suppose the company operating the fleet of self-driving cars has implemented a system so pedestrians and drivers can opt out of data collection by the cars. By going through their online system, people can have images of them blurred so their identity is protected and they cannot be tracked.

35. How many minutes would you spend in the system to successfully opt out?
◯ 0
◯ 1-5
◯ 6-10
◯ 11-15
◯ 16-20
◯ 21-25
◯ 26-30
◯ More than 30

Questions about Uber

36. I feel that companies operating networked fleets of self-driving cars have my best interests in mind
◯ Strongly disagree
◯ Disagree
◯ Neither agree nor disagree
◯ Agree
◯ Strongly agree

37. I feel that Uber's self-driving car division has my best interests in mind
◯ Strongly disagree
◯ Disagree
◯ Neither agree nor disagree
◯ Agree
◯ Strongly agree

38. I feel that Uber has my best interests in mind
◯ Strongly disagree
◯ Disagree
◯ Neither agree nor disagree
◯ Agree
◯ Strongly agree

39. I would have answered the survey questions differently had Uber not been used as the example
◯ Strongly disagree
◯ Disagree
◯ Neither agree nor disagree
◯ Agree
◯ Strongly agree

40. I would trust a different networked self-driving car fleet over Uber's to have my best interests in mind
◯ Strongly disagree
◯ Disagree
◯ Neither agree nor disagree
◯ Agree
◯ Strongly agree

Demographic Questions

41. Please specify your gender
◯ Man
◯ Woman
◯ Other (please specify):
◯ Prefer not to answer

42. Please indicate your age

[textbox]

43. Select the highest education level you have achieved
◯ No high school
◯ Some high school
◯ High school graduate, diploma, or the equivalent
◯ Trade, technical, or vocational training
◯ Some college
◯ Associate degree
◯ Bachelor's degree
◯ Master's degree
◯ Professional or doctoral degree
◯ Prefer not to answer

44. Select the industry in which you work
◯ Accounting
◯ Advertising
◯ Aerospace / Aviation / Automotive
◯ Agriculture / Forestry / Fishing
◯ Biotechnology
◯ Business / Professional Services
◯ Business Services (Hotels, Lodging Places)
◯ Computers (Hardware, Desktop Software)
◯ Communications
◯ Construction / Home Improvement
◯ Consulting
◯ Education
◯ Engineering / Architecture
◯ Entertainment / Recreation
◯ Finance / Banking / Insurance
◯ Food Service
◯ Government / Military
◯ Healthcare / Medical
◯ Internet
◯ Legal
◯ Manufacturing
◯ Marketing / Market Research / Public Relations
◯ Media / Printing / Publishing
◯ Mining
◯ Non-Profit
◯ Pharmaceutical / Chemical
◯ Research / Science
◯ Real Estate
◯ Retail
◯ Telecommunications
◯ Transportation / Distribution
◯ Utilities
◯ Wholesale
◯ Other - Write In
◯ Not applicable

45. Check all that apply:
☐ I have downloaded and used the Uber and/or Lyft mobile apps
☐ I have been a passenger in an Uber and/or Lyft car
☐ I have read an article, viewed a program, or learned online about Uber self-driving cars
☐ I have seen an Uber self-driving car while I was a driver
☐ I have seen an Uber self-driving car while I was a pedestrian
☐ I have ridden in an Uber self-driving car as a passenger
☐ I have not yet ridden, but would ride as a passenger in an Uber selfdriving car
☐ I am or have been an Uber and/or Lyft driver
☐ I have ridden as a safety driver in an Uber self-driving car
☐ I am currently or have previously been employed by Uber directly (not as a driver)
☐ I disagree with actions Uber has taken
☐ I have deleted the Uber app
☐ I have attended a protest against Uber
☐ None of the above

Privacy and Technology Questions

46. Are you majoring in or have a degree or job in computer science, computer engineering, information technology, or a related field?
◯ Yes
◯ No

47. Privacy is really a matter of people's right to exercise control and autonomy over decisions about how their information is collected, used, and shared.
◯ Strongly agree
◯ Agree
◯ Neither agree nor disagree
◯ Disagree
◯ Strongly disagree

48. Control of personal information lies at the heart of privacy.
◯ Strongly agree
◯ Agree
◯ Neither agree nor disagree
◯ Disagree
◯ Strongly disagree

49. I believe that privacy is invaded when control is lost or unwillingly reduced as a result of a marketing transaction.
◯ Strongly agree
◯ Agree
◯ Neither agree nor disagree
◯ Disagree
◯ Strongly disagree

50. Companies seeking information should disclose the way the data are collected, processed, and used.
◯ Strongly agree
◯ Agree
◯ Neither agree nor disagree
◯ Disagree
◯ Strongly disagree

51. It is very important to me that I am aware and knowledgeable about how my personal information will be used.
◯ Strongly agree
◯ Agree
◯ Neither agree nor disagree
◯ Disagree
◯ Strongly disagree

52. It usually bothers me when companies ask me for personal information.
◯ Strongly agree
◯ Agree
◯ Neither agree nor disagree
◯ Disagree
◯ Strongly disagree

53. When companies ask me for personal information, I sometimes think twice before providing it.
◯ Strongly agree
◯ Agree
◯ Neither agree nor disagree
◯ Disagree
◯ Strongly disagree

54. It bothers me to give personal information to so many companies.
◯ Strongly agree
◯ Agree
◯ Neither agree nor disagree
◯ Disagree
◯ Strongly disagree

55. I'm concerned that companies are collecting too much personal information about me.
◯ Strongly agree
◯ Agree
◯ Neither agree nor disagree
◯ Disagree
◯ Strongly disagree

56. To be entered into the raffle for Amazon gift cards, please provide your email address: (We will never use this for purposes out of this research)

[textbox]

57. Is there anything else you would like to add about networked self-driving cars or this survey in general?

[textbox]

Hello,

We would like to thank you again for participating in our study. If you are selected for the raffle, an Amazon gift card code will be sent to the email you provided.

This study is aimed at determining people's awareness and preferences toward the privacy considerations surrounding Uber's self-driving cars. The data you provided will be used to help determine future areas of study and help craft recommendations for the industry in addressing consumer privacy needs and concerns.

Deployed fleets of autonomous vehicles like Uber's self-driving cars are a new phenomenon, and researching these cars in ordinary, real-world scenarios has just begun. From what we know, Uber self-driving cars have three different types of sensors:

1. Radar sensors that map the physical world around the car. They do not collect video and do not store any information; they are just used for navigational purposes.
2. The large camera lens on the roof is used to detect colors, such as those on a traffic light or a stop sign. It does not collect photo or video.
3. Twenty other cameras are used to detect braking vehicles, pedestrians, and other obstacles. Some cameras store video that can be reviewed later manually by people, or via automated computer algorithms.

Some participants in this study were exposed to this information during the study, while others were not. This was done to gauge how people perceive the privacy concerns surrounding Uber's cars with and without context.

Thanks again for your time and ongoing participation in our study. For any further feedback on the study, feel free to email at: selfdrivingcarresearch@cmu.edu

Format vs. Content: The Impact of Risk and Presentation on Disclosure Decisions

Sonam Samat Alessandro Acquisti

Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15232
{sonamsam, acquisti}@andrew.cmu.edu

ABSTRACT

Although the importance of format and presentation of privacy notices has been extensively studied in the privacy literature, less explored is the interplay of presentation and content in influencing users' disclosure decisions. In two experiments, we manipulate the content as well as the format of privacy notices shown to participants who were asked to choose whether they would like to disclose personal information. We manipulate content by changing the objective privacy risk that participants face from disclosing personal information. We manipulate format by changing the manner in which these notices are presented. We find that participants are significantly less likely to share their personal information when the privacy notice is presented under a 'Prohibit [disclosure]' frame, as compared to an 'Allow [disclosure]' frame. However, and importantly, we find that the effect of changes in framing on disclosure decisions is small when the objective privacy risk from disclosure is low, but the effect of framing becomes larger when the risk is increased—that is, for potentially more sensitive decisions. Our results highlight the nuanced interaction effects between the objective content of privacy notices and the manner in which they are presented, on disclosure behavior.

1. INTRODUCTION

Online companies collect many different types of information about their users, such as browsing behavior, search queries, purchase history, location information, and demographic information. Typically, users agree to share this information when they register for an online service and accept the service's privacy policy. In some cases, the permission to collect specific types of information is obtained after the registration process is complete, while the user is using the service. For example, a mobile app may obtain consent for collecting browsing behavior and purchase history in its privacy policy, but may later display a prompt asking for permission to collect location information while the user is employing a feature that specifically requires the use of location information. In both cases, the service provider designs the interface where the user makes his or her choice to disclose personal information—thus, it can act as a "choice architect" [21], and influence users' decisions and behaviors. Substantial behavioral research in the privacy field has, in fact, suggested that the interface itself, and not just the content of the policy, may affect individuals' propensity to disclose personal information [e.g., 2]. Much less studied, however, is how the effect of changes in the presentation of privacy-relevant information interacts with the effect of changes in the objective privacy risk from disclosure, on individuals' propensity to disclose personal information.

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12 -- 14, 2017, Santa Clara, California.

In two experiments, we manipulate the content as well as the format of privacy notices shown to participants. We manipulate content by changing the privacy risk that participants face from disclosing information (for example, by varying the entity with which the information is to be shared). We manipulate format by changing the frame under which these notices are presented to the subjects. We find that participants are significantly less likely to share their personal information when the privacy notice is presented under a 'Prohibit [disclosure]' frame, as compared to an 'Allow [disclosure]' frame. However, and importantly, we also find that the effect is small when the objective privacy risk from disclosure is low, but becomes larger when the risk is increased to moderate levels. The results highlight the nuanced interactions between the actual content of privacy notices and the way they are presented, in influencing consumer behavior.

The implications of the results are twofold. First, these results highlight the challenges of relying solely on providing notice and choice to consumers to achieve a policy maker's goal of consumer privacy protection. The manner in which notices are framed can have a significant effect on behavior. As long as firms are the choice architects of their own privacy notices, they may implement framing nudges that influence consumers' choices, for instance to affect the rate of disclosure of personal information. Second, these results provide insights into identifying specific situations where framing effects matter the most (when objective risks are moderate), thus helping organizations, individuals, and policy makers direct their attention to notices that may lead to strong framing effects.

2. THEORETICAL BACKGROUND

Framing effects refer to the phenomena whereby "simple and unspectacular changes" in the presentation of decision problems lead to changes in choice [11]. These simple changes do not alter the objective factors of the decision. Evidence from behavioral decision research shows that such seemingly insignificant changes can have a significant impact on individuals' choices. In other words, they can act as "nudges." In 1981, Tversky and Kahneman presented participants with the choice between a certain treatment that can save 200 of 600 people affected by a disease, and a probabilistic option that has the same expected value (1/3rd probability that 600 people will be saved). They found that participants prefer the certain choice to the probabilistic one, but when the same choice is framed in terms of the number of lives lost (i.e., 400 people die out of 600 affected) then participants preferred the probabilistic option to the certain choice [22]. The mere framing of the choices (in terms of lives saved or lost) had a significant impact on individuals' decisions. In 1988, Levin and Gaeth showed that labeling a pack of ground beef as "75% lean" instead of "25% fat" significantly impacted participants' perception of the quality of the beef [14]. Along similar lines, other researchers have found differences in perceptions when situations are described in terms of success rates versus failure rates [6].

Even though framing effects have been shown to significantly impact behavior in many cases, there is also no dearth of examples where framing effects have failed to change behavior ([15] provides a review of when framing effects have been shown to appear and disappear). For example, Druckman's 2001 work showed that framing effects can drastically decrease and even diminish when individuals are provided with credible advice about how to make the decision [7]. The intensity of framing effects also changes dramatically across different task domains. In 1996, Wang studied framing effects in risky choices across three different task domains, where participants had to choose between a sure outcome and a gamble of the same expected value. In addition to finding significantly different intensities of framing effects across domains, he also found that, within a given domain, framing effects tend to appear and disappear depending on the expected value of the gamble [23]. Since expected value manipulations change the objective benefits provided, this result suggests that when objective content is varied framing effects may change in intensity and may even disappear.

In the typical framing studies, participants are asked to choose between a sure option and a gambling option, while the framing of these options is manipulated between conditions. Under the positive frame, participants choose between winning an amount X for sure and a gamble in which they may win an amount greater than X but with some probability less than 1 (for example, a gamble to win 2X with probability ½). Under the negative frame, participants are given some money in advance and asked to choose between giving back an amount X for sure and a gamble in which they may lose an amount greater than X but with some probability less than 1 (for example, a gamble to lose 2X with probability ½). Participants tend to prefer the sure option to the probabilistic one when choices are framed as gains, and the opposite occurs when choices are framed as losses. But some researchers have found that framing effects can disappear if the payoffs from the choices are too small [9]. Therefore, it is important to investigate the interplay between objective payoffs and framing. In the case of privacy decisions, framing effects may disappear when the objective risk associated with disclosing personal information is perceived to be too low. In our studies, we test the interplay between objective risk and framing effects by varying both factors: 1) the objective risk participants face when disclosing personal information and 2) the manner in which the risks from disclosure are framed.

Previous researchers have also argued that non-normative factors, such as framing, tend to have an impact on decisions when consumers' preferences are ambiguous, but this effect diminishes (and even disappears) when preferences are more certain [19]. This is because, when preferences are ambiguous, individuals tend to look for additional cues (such as how a problem is framed) to help them construct preferences [18, 20]. This suggests that framing may significantly impact decisions when individuals' preferences are ambiguous, and may fail to impact decisions when preferences are certain. In our work, we test this conjecture by measuring participants' level of ambiguity or uncertainty with their sharing decisions, and evaluating whether uncertainty can explain how framing effects impact disclosure decisions.

3. EXPERIMENTAL APPROACH

In two studies, we manipulate the manner in which privacy notices that inform participants about the risks from disclosure are framed. These studies were conducted on Amazon's Mechanical Turk (MTurk) platform. Previous researchers have shown that MTurk workers are more demographically diverse than the typical convenience samples of American college students, and established results have been replicated with this population, confirming its reliability [3, 4]. Amazon Mechanical Turk also allows researchers to approve or reject participants' payment based on their performance. Therefore, each participant has an approval rating, which is the percentage of his or her previously completed surveys or tasks that have been approved. We implemented a minimum requirement of a 95% approval rating during our recruitment process, and also used attention check questions in our surveys to ensure high quality data.

Our framing manipulation was embedded in the choice presented to our participants. In one condition, we asked participants if they would like to "allow us to share your information" and in another we asked if they would like to "prohibit us from sharing your information." Similar to framing manipulations used in previous literature, this manipulation allowed us to change the format of the notice while keeping the objective content of the notices constant. Query Theory research [8, 10] suggests that under the 'Allow' frame individuals will be more likely to give permission to share their information because 'Allow' frames typically make individual more likely to think about reasons to act as described in the question (in our context, allow the sharing of their information); whereas individuals will be less likely to give permission to share their information under the 'Prohibit' frame because it makes them more likely to think about reasons to prohibit the sharing of their information. Therefore, we hypothesize that participants in the 'Allow' condition will be more likely to accept the privacy policy than those in the 'Prohibit' condition.

Based on the literature highlighted in the previous section, we expect that, at low levels of objective privacy risk, individuals' sharing decisions will not vary with the framing of the notice. On the other hand, at moderate levels of disclosure risk, framing will have a significant effect on individuals' sharing decisions. We tested this hypothesis by varying the objective risk from disclosing personal information between conditions, and comparing the impact of framing at the different objective risk levels. In addition, we also tested the conjecture that framing effects depend on the level of ambiguity or uncertainty that participants have towards these disclosure decisions. Specifically, we investigated whether framing effects only tend to appear when participants are less certain about their preferences. All our studies were approved by the Institutional Review Board at Carnegie Mellon University. The IRB review covered the deception that we implemented in our studies. All participants were debriefed at the end of the study to clarify that their data will not be shared with anyone other than the researchers conducting this study.

4. STUDY 1

4.1 Methods

In this study, we manipulated risk and framing in a context that involved making real information disclosure decisions. Participants were recruited from Amazon's Mechanical Turk for a study about ethical behaviors. Following the design put forward by Adjerid et al. in their 2014 work, we first asked participants for their demographic information, and informed them that they would be asked several questions of a sensitive nature, such as "Have you ever had a one-night stand?" The goal of collecting demographic information before presenting the disclosure choice was to elicit a level of quasi-identifiability, so the subsequent disclosure decisions would not seem entirely risk-free [2]. Then we asked participants to make a disclosure choice for whether they would be willing to share their responses to the ethical behavior questions with a specific audience. We manipulated objective privacy risk by changing this audience: in one condition, the audience was research assistants for this study, whereas in the other condition the audience was a marketing company. We expected that participants would perceive sharing information with research assistants as not very risky, but sharing with a marketing company would be perceived as being somewhat risky. Framing was manipulated using the 'Allow' and 'Prohibit' frames. For instance, participants in the 'Marketing Company' and 'Allow' frame condition were shown the following sentence: "Allow my responses to be shared with a marketing company." As participants were aware that they were going to be asked a set of highly intrusive questions, the decision to share their responses with the specified audience involved evaluating actual risks, as opposed to hypothetical ones. Following this question, we asked them ten questions related to ethically questionable activities.[1] Participants were informed that if they were not comfortable answering any of these questions, they could skip them and proceed with the survey without any penalty. Note that while participants were told that their information would be shared with the specified audience, we did not actually share their information with anyone outside of the primary researchers associated with this study. Participants were debriefed about our real motivations at the end of the study.

4.2 Results

Three hundred and seventy-six individuals (Mean Age = 32.5; 58% Male) from Amazon's Mechanical Turk participated in Study 1. Participants in the 'Allow' condition were 21% more likely to share their responses compared to those in the 'Prohibit' condition (86% vs. 71%, $\chi 2$ (1) = 12.22, p<0.001). Furthermore, participants in the 'Research Assistants' condition were 55% more likely to share their responses when compared to those in the 'Marketing Company' condition (96% vs. 62%, $\chi 2$ (1) = 64.74, p<0.001). Therefore, we observe main effects of framing and risk.

Looking at the two risk conditions individually, we find that there is no significant effect of framing among participants in the 'Research Assistants' condition (98% vs. 93%, $\chi 2$ (1) = 2.27, p=0.132) but the effect is significant in the 'Marketing Company' condition (74% vs. 49%, $\chi 2$(1) = 11.77, p=0.001). In other words, at the level of risk presented in the 'Research Assistants' condition, almost everyone is willing to share his or her responses irrespective of framing. Therefore, we find that when the objective privacy risk is low, our framing manipulation does not significantly impact disclosure rates.

Next, we test the following econometric model:

$$Share_i = \beta_0 + \beta_{ResearchAssistants}\ ResearchAssistants + \beta_{Allow}\ Allow + \beta_{ResearchAssistants*Allow}\ ResearchAssistants*Allow + \varepsilon_i$$

where 'Share' represents whether participants gave permission to share their data or not (binary choice), 'ResearchAssistants' is a dummy variable equal to 1 for the 'Research Assistants' condition, 'Allow' is a dummy variable equal to 1 for the 'Allow' condition, 'ResearchAssistants*Allow' is the interaction term, and 'ε' is the random error term. Our dependent variable is a binary choice so we estimate the model as a probit. Since interaction terms are easier to interpret with linear regression coefficients, we report the OLS regression coefficients here. Probit model results are consistent with the OLS results. Both probit and OLS complete results are reported in Appendix B.

Estimation of the model without the interaction term confirms the main effects of framing and objective risk (β_{Allow} = 0.145, p<0.001; $\beta_{ResearchAssistants}$ = 0.340, p<0.001). The model with the interaction term also shows a significant and positive coefficient on 'Allow' (β_{Allow} = 0.242, p<0.001), indicating that framing has a significant effect on sharing decisions in the 'Marketing Company' condition. A significant and positive coefficient on 'Research Assistants' ($\beta_{ResearchAssistants}$ = 0.440, p<0.001) indicates that the level of objective privacy risk affects decisions to share in the 'Prohibit' condition. The interaction term in this model is also statistically significant ($\beta_{ResearchAssistants*Allow}$ = –0.197, p=0.009), confirming that the intensity of the framing effect is significantly smaller in the 'Research Assistants' condition than in the 'Marketing Company' condition. Figure 1 shows these effects.

This interaction effect provides evidence for the argument that the intensity of the framing effect depends on the level of risk that participants are faced with. When the risk of information sharing is low (such as sharing survey responses with research assistants) then framing effects may disappear, but they appear when the risk of information sharing is relatively higher (such as sharing survey responses with a marketing company). In addition, these results also suggest that an increase in objective privacy risk under the 'Allow' frame leads to a smaller adjustment of sharing behavior, as compared to an equivalent increase in risk under the 'Prohibit' frame. This implies that, when privacy policies are framed in a positive way (as most current day privacy policies are), individuals may be less likely to adjust their sharing behavior to account for an increase in objective privacy risk, compared to cases when privacy policies are framed in a negative way. This result is important as companies frequently make changes to their privacy policies and

[1] We used questions rated as highly intrusive in Acquisti et al.'s 2012 work [1]. See Appendix A.

these changes often involve increasing the privacy risk for consumers.

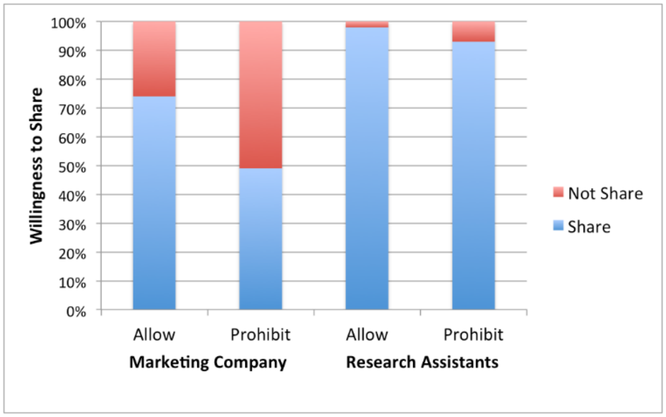


Figure 1. Percentages of participants willing to share their information in Study 1.

4.3 Discussion

The results from Study 1 provide evidence that framing nudges in privacy notices can influence disclosure decisions, but these effects are not universal across different levels of objective privacy risk. Positive frames such as 'Allow' make participants more likely to share their personal information when compared to negative frames such as 'Prohibit'. But there may be credible levels of objective privacy risk at which individuals overcome the effect of framing, because the risk associated with sharing is perceived to be too low (such as sharing survey responses with research assistants). When the objective privacy risk associated with sharing information is moderate (such as sharing with a marketing company), framing effects are more likely to occur. This result is consistent with previous research that showed framing effects tend to disappear when objective payoffs from a gamble are small [9, 12, 13].

5. PRELIMINARY ANALYSIS

Following Study 1, we conducted a set of additional studies to measure the level of perceived risk associated with sharing personal information with different entities. These studies gave us insight into the relative perceptions of risk associated with the two conditions tested in Study 1 as well as other conditions which we later tested in Study 2.

5.1 Survey A

Participants were recruited from Mechanical Turk for a study in which they would be asked for their opinions about a hypothetical scenario. Participants were asked to imagine that they were answering a survey about ethical behaviors on Mechanical Turk, which involved answering sensitive questions such as "Have you ever had a one-night stand?" They were then told that the researchers of that study first asked them whether they would be willing to share their information with a specific audience. We manipulated this audience between conditions, testing four different audiences: 1) research assistants, 2) marketing companies, 3) publicly on the Internet, and 4) on Mechanical Turk forums (such as Turker Nation, MTurk Grind, MTurk Forum, etc.) along with their Mechanical Turk ID. The last condition was included in an attempt to increase the personal relevance of the decision for Mechanical Turk participants, as individuals on Mechanical Turk often use such forums to discuss pragmatic considerations about MTurk tasks such as pay rates or requesters' reputations [5]. Arguably, they may care a significant amount about their reputation on these forums. As Mechanical Turk does not permit the collection of any personally identifiable information, we attempted to achieve quasi-identifiability by claiming that the responses would be shared along with participants' Mechanical Turk ID. Next, participants were asked four questions, each on a 1–7 scale from 'Not at all' to 'Very much': 1) how risky they thought it would be to share their survey responses with this audience, 2) how likely they would be to share their survey responses with this audience, 3) how comfortable they would be sharing their survey responses with this audience, and 4) how concerned they would be about sharing their survey responses with this audience.

5.1.1 Results

One thousand, two hundred seventy-nine participants (Mean Age = 32.3; 55% Male) from Amazon Mechanical Turk completed this survey. As shown by the mean values reported in Table 1, participants did not think sharing their survey responses with research assistants was very risky. Sharing survey responses with marketing companies, publicly online, and on Mechanical Turk forums with Mechanical Turk IDs were perceived to be increasingly riskier scenarios, in that order.[2] These results confirm the assumption made in Study 1, that the perceived risk of sharing survey responses with research assistant is very low risk while that of sharing survey responses with marketing companies is moderate.

Table 1. Mean values of variables measured in Survey A.

Dependent Variable	Research Assistants	Marketing Companies	Publicly Online	On MTurk with Turk ID
Risky	2.46	3.75	4.32	4.44
Likely	5.67	3.98	3.12	2.88
Comfortable	5.37	4.11	3.32	3.22
Concerned	2.60	3.84	4.56	4.53

[2] The differences between 'Publicly Online' and 'On MTurk with Turk ID' for all four variables are not statistically significant. All other differences are statistically significant at the 0.001 level.

5.2 Survey B

Participants were recruited for a survey about information sharing preferences. Each participant was shown five scenarios, two of which involved sharing information with a hypothetical news website and three of which involved sharing survey responses to ethical behavior questions with a specific audience. Therefore, this study was conducted with a within-subjects design as opposed to the between-subjects design of Survey A. In the two scenarios involving the sharing of information with a hypothetical news website, participants were asked to imagine that they want to read an article on a news website but are faced with the decision to accept or reject the news website's privacy policy before they can read the article. We tested participants' opinions about two different amounts of personal information purportedly being collected by the news website. (These two scenarios are used in studies that are not reported in this paper.) The scenarios involving sharing survey responses with specific audiences were presented in the same way as in Survey A (by asking participants to imagine they are answering a survey about ethical behaviors). The three audiences tested in this survey were: 1) research assistants, 2) other research organizations, and 3) marketing companies. We asked participants how risky they thought it would be to share the information and how likely they would be to share it, on a 1–7 scale ranging from 'Not at all' to 'Very much.' The same two questions were used for the hypothetical news website scenarios as well, and the order of these two questions was randomized across participants.

5.2.1 Results

One hundred twenty participants (Mean Age = 34.3; 67% Male) from Amazon Mechanical Turk completed this study. As shown by the mean values reported in Table 2, sharing survey responses with research assistants, other research organizations, and marketing companies are perceived to be increasingly riskier scenarios, in that order.[3]

Table 2. Mean values of variables measured in Survey B.

Dependent Variable	Research Assistants	Other Research Organizations	Marketing Companies
Risky	2.53	3.24	4.03
Likely	5.43	4.70	3.56

6. STUDY 2

6.1 Methods

Based on the results of the surveys we designed the second study. In Study 2, we test framing effects at three different risk levels to get a better sense of how framing effects vary with risk. The design of this study is similar to that of Study 1, but instead of a 2-by-2 design, we used a 2-by-3 design. Two objective privacy risk levels tested in this study are similar to the ones used Study 1: sharing with research assistants and sharing with marketing companies. The third risk level is sharing on Mechanical Turk forums (such as Turker Nation, MTurk Grind, MTurk Forum, etc.), along with their Mechanical Turk ID. The framing manipulation is implemented in the same way as in Study 1, by altering whether participants are asked to 'Allow' the sharing of their information or 'Prohibit' the sharing of their information. Just as in Study 1, participants were first informed that they were going to be asked sensitive questions. Then, they were asked for their sharing preferences (varying the scenarios in their framing and risk between conditions) and were subsequently asked the same ethical questions as used in the previous study. We include an additional question in this study that measures participants' level of uncertainty with the sharing decision, using the uncertainty subscale from the Decision Conflict Scale [16]. This question is included to test whether participants' level of uncertainty with the sharing decision varies in the same way as framing effects vary when objective risk is manipulated (i.e., more uncertainty correlates with larger framing effects). Such a result would provide support for the conjecture that framing effects only occur when individuals' preferences are ambiguous, consistent with previous research [19]. Participants were debriefed at the end of the study.

6.2 Results

Nine hundred ninety-five individuals (Mean Age = 37.2; 47% Male) from Amazon Mechanical Turk completed Study 2. Participants in the 'Allow' condition were 43% more likely to share their responses when compared to those in the 'Prohibit' condition (73% vs. 51%, $\chi 2$ (1) = 51.95, $p<0.001$). Furthermore, participants in the 'Research Assistants' condition were 64% more likely to share their responses when compared to those in the 'Marketing Company' condition (92% vs. 56%, $\chi 2$ (1) = 109.26, $p<0.001$), and participants in the 'Marketing Company' condition were 37% more likely to share their responses when compared to those in the 'Mechanical Turk Forums' condition (56% vs. 41%, $\chi 2$ (1) = 14.64, $p<0.001$). Therefore, we observe the main effects of framing and risk.

Looking at the three objective privacy risk conditions individually, we find a significant effect of framing at all three risk levels ('Research Assistants' condition: 96% vs. 87%, $\chi 2$ (1) = 7.84, $p=0.005$; 'Marketing Company' condition: 67% vs. 44%, $\chi 2$ (1) = 17.42, $p<0.001$; 'Mechanical Turk Forums' condition: 58% vs. 22%, $\chi 2$ (1) = 45.11, $p<0.001$). The relative size of the framing effect increases as risk increases (Cohen's d for 'Research Assistants' condition = 0.31; Cohen's d for 'Marketing Company' condition = 0.47; Cohen's d for 'Mechanical Turk Forums' condition = 0.78). Therefore, we find that when objective risk from disclosure is low, our framing manipulation has a smaller impact on sharing decisions, but as the objective risk increases, the effect of our framing manipulation also increases. It is important to note that even the highest level of risk tested in this study ('Mechanical Turk Forums') is perceived to be 'moderate', as shown by the mean value of perceived risk in Survey A (mean perceived risk for 'Mechanical Turk Forums' = 4.44 on a 1–7 scale). So, the increasing trend in the size of framing effects is observed when risk increases from low to moderate levels. We do not know how framing effects may vary when the perceived risk increases beyond moderate levels to high levels.

[3] All differences are statistically significant at the 0.001 level.

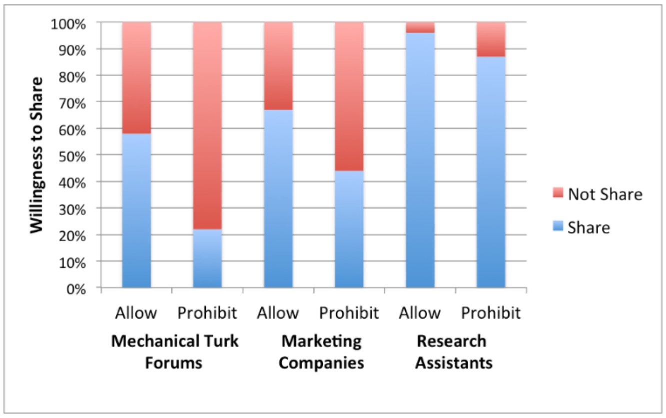


Figure 2. Percentages of participants willing to share their information in Study 2.

Estimation of the model as in Study 1 without the interaction term (coding the risk variable as '0' for 'Research Assistants', '1' for 'Marketing Companies', and '2' for 'Mechanical Turk Forums') confirms the main effects of framing and objective risk (β_{Allow} = 0.226, p<0.001; β_{Risk} = –0.255, p<0.001). The model with the interaction term also shows a significant and positive coefficient on 'Allow' (β_{Allow} = 0.087, p=0.039), indicating that framing has a significant effect on sharing decisions in the 'Research Assistants' condition. A significant and negative coefficient on 'Risk' (β_{Risk} = –0.325, p<0.001) indicates that the level of risk affects decisions to share in the 'Prohibit' condition (increasing risk decreases willingness to share). The interaction term in this model is also statistically significant ($\beta_{Risk*Allow}$ = 0.137, p<0.001), confirming that the intensity of the framing effect significantly changes as risk is varied. The positive coefficient on the interaction term confirms that the size of the framing effect increases as the level of risk increases. Figure 2 shows these effects graphically.

Next, we analyze how participants' uncertainty varies with objective risk, to see if uncertainty can explain the increase in the size of framing effects as risk increases. The three items in the uncertainty subscale of the Decisional Conflict Scale [16] show high reliability (Cronbach's alpha = 0.927) so we follow the instructions provided by O'Connor [17] to code the 1–5 scale (from 'strongly agree' to 'strongly disagree') as 0–4, then average the three items and multiply the averaged value by 25. The final score ranges from 0 (participant feels extremely certain about the best choice for them) to 100 (participant feels extremely uncertain about the best choice for them). The uncertainty values do not vary significantly with risk (β_{Risk} = 0.318, p=0.715). The mean uncertainty values across the three risk conditions do not show an increasing trend (mean uncertainty value for 'Research Assistants' condition = 21.67; mean uncertainty value for 'Marketing Companies' condition = 24.47; mean uncertainty value for 'Mechanical Turk Forums' condition = 22.33). This suggests that the level of uncertainty, as measured by the uncertainty subscale of the Decisional Conflict Scale, cannot explain the increase in size of framing effects as risk increases.

6.3 Discussion

The results from Study 2 confirm an increasing trend in the size of framing effects as objective privacy risk increases. We find a small but significant effect of framing on sharing decisions when perceived risk is low (such as sharing with research assistants). The size of the framing effect increases as risk is increased to moderate levels (such as sharing on Mechanical Turk Forums along with Mechanical Turk ID).

These results are particularly important from a policy perspective as they suggest that regulators should work towards more nuanced requirements in terms of how privacy notices ought to be framed. Individuals' propensity to framing effects varies considerably with objective risk. A single set of blanket requirements for all websites (irrespective of the amount of privacy risk consumers face from sharing information) may not be sufficient to protect consumers' privacy. For instance, websites that merely collect users' IP addresses versus those that collect web browsing and purchase behaviors should not be subject to the same regulations. Our results suggest that the latter category of websites may be more capable of nudging consumers' sharing decisions by using framing nudges, and therefore should be held to higher standards by policy makers. While it is challenging to present information in a truly "neutral" frame, further work should investigate solutions to mitigate the privacy risk faced by consumers, especially in situations with moderate objective risk.

It is important to discuss the limitations of this study. First, the highest level of objective privacy risk tested in this study (sharing on Mechanical Turk Forums along with Mechanical Turk ID) is only perceived to be moderately risky by our participants. So, the observed increasing trend in framing effects can only be claimed to occur when risk increases from low to moderate levels. While we do not know for sure how framing effects would vary when risk is increased from moderate to high levels, we suspect that framing effects may decrease at very high levels of risk. Second, this study also attempted to test whether participants' level of uncertainty can explain the increase in size of framing effects as risk is increased. We could not find evidence for this, as uncertainty did not vary significantly across the different risk levels. Further research is required to better understand why framing effects increase when objective risk is increased from low to moderate levels.

7. CONCLUSION

In two experiments, we studied the influence of a framing nudge that may be used in privacy notices to influence individuals' willingness to disclose personal information. We found that the manner in which privacy notices are framed has a significant impact on individuals' disclosure decisions. Consistent with our hypotheses, notices that use the 'Prohibit' frame reduce the likelihood that individuals will share their information as compared to notices that use the 'Allow' frame. However, and importantly, we also found that the intensity of this framing effect is small when objective privacy risk from disclosure is small, and it increases as the objective privacy risk increases to moderate levels.

These findings have implications for the design of privacy policies that can empower consumers to face the tradeoffs between privacy risks and the benefits associated with data sharing. Specifically, our results strengthen the notion that simply providing consumers with notice and choice may not be sufficient mechanisms to serve the goal of consumer privacy protection. The manner in which notice and choice are framed is also important as companies may use framing nudges to impact individuals' sharing decision. Our results also assist in guiding the attention of organizations and policy makers towards cases where individuals might be most susceptible to framing nudges, specifically when the objective privacy risks are moderate. Owing to the substantially different intensities of framing effects observed at different objective privacy risk levels, it appears that a blanket policy for all websites, irrespective of the amount of privacy risk consumers face from sharing information with the site, may not be sufficient to protect consumers' privacy. Companies and policy makers may consider more nuanced sets of rules concerning how privacy policies should be framed, keeping in mind the level of privacy risks put forward by different websites.

8. ACKNOWLEDGMENTS

This research has been in part supported by the National Science Foundation under grants 1012763, 1514192, and 1327992. We would like to thank all reviewers for their feedback and suggestions.

9. REFERENCES

[1] Acquisti, A., John, L. K., & Loewenstein, G. (2012). "The impact of relative standards on the propensity to disclose," *Journal of Marketing Research, 49(2)*, 160-74.

[2] Adjerid I., Acquisti, A., & Loewenstein, G. (2014). "Framing and the Malleability of Privacy Choices," *In Workshop on the Economics of Information Security (WEIS) 2014*.

[3] Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). "Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk," *In Political Analysis*, *20(3)*, 351-68.

[4] Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). "Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data?" *Perspectives on Psychological Science, 6(1)*, 3-5.

[5] Chandler, J., Mueller, P., & Paolacci, G. (2014). "Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers," *Behavior Research Methods, 46(1)*, 112-30.

[6] Davis, M. A., & Bobko, P. (1986). "Contextual effects on escalation processes in public sector decision making," *Organizational Behavior and Human Decision* Processes, 37, 121-38.

[7] Druckman, J. N. (2001). "Using credible advice to overcome framing effects," *Journal of Law, Economics, and Organization, 17(1)*, 62-82.

[8] Hardisty, D. J., Johnson, E. J., & Weber, E. U. (2010). "A dirty word or a dirty world? Attribute framing, political affiliation, and query theory," *Psychological Science, 21(1)*, 86-92.

[9] Hsee, C. K., & Weber, E. U. (1997). "A fundamental prediction error: Self–others discrepancies in risk preference," *Journal of Experimental Psychology: General, 126(1)*, 45.

[10] Johnson, E. J., Häubl, G., & Keinan, A. (2007). "Aspects of endowment: a query theory of value construction," *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(3)*, 461.

[11] Kühberger, A. (1998). "The influence of framing on risky decisions: A meta-analysis," *Organizational Behavior and Human Decision Processes, 75(1)*, 23-55.

[12] Kühberger, A., Schulte-Mecklenbeck, M., & Perner, J. (1999). "The effects of framing, reflection, probability, and payoff on risk preference in choice tasks," *Organizational Behavior and Human Decision Processes, 78(3)*, 204-31.

[13] Kühberger, A., Schulte-Mecklenbeck, M., & Perner, J. (2002). "Framing decisions: Hypothetical and real," *Organizational Behavior and Human Decision Processes, 89(2)*, 1162- 75.

[14] Levin, I. P., & Gaeth, G. J. (1988). "Framing of attribute information before and after consuming the product," *Journal of Consumer Research, 15*, 374-78.

[15] Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). "All frames are not created equal: A typology and critical analysis of framing effects," *Organizational Behavior and Human Decision Processes, 76(2)*,149-88.

[16] O'Connor, A. M. (1995). "Validation of a decisional conflict scale," *Medical Decision Making, 15(1)*, 25-30.

[17] O'Connor, A. M. (updated 2010). *User Manual – Decisional Conflict Scale*, https://decisionaid.ohri.ca/docs/develop/User_manuals/UM_Decisional_Conflict.pdf

[18] Payne, J. W., Sagara, N., Shu, S. B., Appelt, K. C., & Johnson, E. J. (2013). "Life expectancy as a constructed belief: Evidence of a live-to or die-by framing effect," *Journal of Risk and Uncertainty, 46(1)*, 27-50.

[19] Schoorman, F. D., Mayer, R. C., Douglas, C. A., & Hetrick, C. T. (1994). "Escalation of commitment and the framing effect: An empirical investigation," *Journal of Applied Social Psychology, 24(6)*, 509-28.

[20] Slovic, P. (1995). "The construction of preference," *American Psychologist, 50*, 364-71.

[21] Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving Decisions About Health, Wealth and Happiness*, New Haven and London: Yale University Press.

[22] Tversky, A., & Kahneman, D. (1981). "The framing of decisions and the psychology of choice," *Science, 211(4481)*, 453-58.

[23] Wang, X. T. (1996). "Framing effects: Dynamics and task domains," *Organizational Behavior and Human Decision Processes, 68(2)*, 145-57.

APPENDICES

A. Appendix A

Questions used in Study 2 (from Acquisti et al., 2012):

1. Have you ever had sex with the current husband, wife, or partner of a friend?

2. Have you ever masturbated at work or in a public restroom?

3. Have you ever had a fantasy of doing something terrible (e.g., torturing) to someone?

4. Have you ever fantasized about having violent non-consensual sex with someone?

5. Have you ever, while an adult, had sexual desires for a minor?

6. Have you ever neglected to tell a partner about a sexually transmitted disease from which you were suffering?

7. Have you ever had sex with someone who was too drunk to know what they were doing?

8. Have you ever stolen anything that did not belong to you?

9. Have you ever tried to gain access to someone else's (e.g., a partner, friend, or colleague's) email account?

10. Have you ever looked at pornographic material?

B. Appendix B

OLS regression coefficients for sharing choice in Study 1

	(1)	(2)	(3)
	Share	Share	Share
ResearchAssistants	0.340***	0.439***	0.441***
	(0.038)	(0.058)	(0.058)
Allow	0.145***	0.242***	0.246***
	(0.038)	(0.069)	(0.068)
ResearchAssistants *Allow		–0.197***	–0.194**
		(0.075)	(0.075)
Male			–0.048
			(0.038)
Age			0.004**
			(0.002)
Constant	0.544***	0.495***	0.381***
	(0.041)	(0.052)	(0. 082)

*p<0.10; **p<0.05; ***p<0.01
Standard errors in brackets

Probit results for sharing choice in Study 1

	(1)
	Share
	Average Marginal Effects
ResearchAssistants	0.339***
	(0.037)
Allow	0.143***
	(0.037)
ResearchAssistants *Allow	–0.197***
[Cross-partial derivative]	(0.075)

*p<0.10; **p<0.05; ***p<0.01
Standard errors in brackets

OLS regression coefficients for sharing choice in Study 2

	(1)	(2)	(3)
	Share	Share	Share
Risk	–0.255***	–0.325***	–0.325***
	(0.016)	(0.023)	(0.023)
Allow	0.226***	0.087**	0.092**
	(0.027)	(0.042)	(0.042)
Risk *Allow		0.137***	0.133***
		(0.033)	(0.033)
Male			0.052*
			(0.027)
Age			–0.002*
			(0.001)
Constant	0.765***	0.835***	0.845***
	(0.025)	(0.030)	(0. 040)

*p<0.10; **p<0.05; ***p<0.01
Standard errors in brackets

Probit results for sharing choice in Study 2

	(1)
	Share
	Average Marginal Effects
Risk	–0.240***
	(0.012)
Allow	0.217***
	(0.025)
Risk*Allow	0.098***
[Cross-partial derivative]	(0.042)

*p<0.10; **p<0.05; ***p<0.01
Standard errors in brackets

New Me: Understanding Expert and Non-Expert Perceptions and Usage of the Tor Anonymity Network

Kevin Gallagher
New York University
2 Metrotech Center
Brooklyn, NY 11201
kevin.gallagher@nyu.edu

Sameer Patil
Indiana University &
New York University
901 E. 10th Street
Bloomington, IN 47408
patil@indiana.edu

Nasir Memon
New York University
2 Metrotech Center
Brooklyn, NY 11201
memon@nyu.edu

ABSTRACT

Proper use of an anonymity system requires adequate understanding of how it functions. Yet, there is surprisingly little research that looks into user understanding and usage of anonymity software. Improper use stemming from a lack of sufficient knowledge of the system has the potential to lead to deanonymization, which may hold severe personal consequences for the user. We report on the understanding and the use of the Tor anonymity system. Via semi-structured interviews with 17 individuals (6 experts and 11 non-experts) we found that experts and non-experts view, understand, and use Tor in notably different ways. Moreover, both groups exhibit behavior as well as gaps in understanding that could potentially compromise anonymity. Based on these findings, we provide several suggestions for improving the user experience of Tor to facilitate better user understanding of its operation, threat model, and limitations.

1. INTRODUCTION

In the past, it was often sufficient to exclude one's name from an interaction to protect one's identity. The information age, however, requires more advanced means to achieve anonymity [21]. Many anonymity systems have risen to meet that demand. These anonymity systems play a vital role in the lives of society's important actors, such as journalists, activists, dissidents, law enforcement agents, and individuals for whom the disclosure of identity could lead to severe consequences. Further, these systems provide a means to assert important civil liberties, such as privacy, freedom of expression, etc. In fact, the use of such tools by the general population has experienced a large rise[1] in the aftermath of Edward Snowden's revelations of mass surveillance activities of the National Security Agency (NSA). However, incorrect use of these systems can lead to deanonymization which, in turn, can lead to a variety of consequences, ranging from slight embarrassment to imprisonment and, in extreme circumstances, death. Additionally, the strength of the anonymity system depends on the number of indistinguishable users [3]. As a result, when a user deanonymizes him- or herself, he or she weakens the strength of the anonymity network as a whole. Proper understanding and use of anonymity tools, therefore, play an important role in ensuring accurate and effective achievement of anonymity via the system. Yet, there has been surprisingly little research that looks into how users understand and conceptualize the underlying operation of these systems. We aim to address this gap.

Though many anonymity systems with different technical details and threat models [2, 34] exist, currently the most popular anonymity system is Tor [4]. Tor is a low-latency network that provides anonymity when performing tasks such as Web browsing. It works using the concept of onion routing [8], which routes traffic through multiple volunteer-run nodes, removing a layer of encryption at each node. By default, the number of nodes is set to three, which is the minimum number of nodes required to achieve anonymity. With a circuit of three or more hops, each node knows the identities of only the immediate predecessor and successor. As a result, no node knows both the source and the destination of a message. When the traffic arrives at the last node, or the 'exit node,' the plain text of the message is forwarded to the destination. In addition, Tor provides support for Onion Services,[2] which allow a server and a client to contact each other without knowing each other's IP addresses.

Tor is used globally by a wide variety of people. According to estimates by the Tor Project, Tor averages around 1,750,000–2,000,000 unique users each day from all around the world.[1] Many of these individuals are located in nations with oppressive regimes. For instance, recent estimates of Tor usage show that between June and August 2016, daily number of users of Tor in Iran ranged from 10,500–12,000.[1] Individuals based in such countries use Tor to access information from sources forbidden or censored by their nations and to pass along information about abuses of their governments to parties who can publish it without fear of retribution. Additionally, the Tor network provides the underlying platform for chat programs, such as ricochet,[3] and file shar-

[1] https://metrics.torproject.org

[2] Onion Services were previously referred to as Hidden Services, a term still used by some users.

[3] https://ricochet.im

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

ing programs, such as SecureDrop,[4] that allow journalists to communicate with sources confidentially and anonymously. As mentioned earlier, Tor is used by ordinary citizens seeking to escape ubiquitous surveillance [26] and censorship. Therefore, inaccurate use of Tor that leads to deanonymization holds potential for great individual and societal harm.

Due to its popularity and importance as an anonymity tool, we focused on Tor to investigate people's use of anonymity systems along with their understanding of the threat model and system operation. Specifically, we addressed the following research questions:

1. Why do people use Tor?
2. How well do users understand the underlying operation of the Tor system?

We tackled the above research questions by conducting semi-structured interviews with a diverse sample of 17 Tor users. Based on an analysis of the interview responses, we make the following contributions:

- We describe user perceptions and practices regarding Tor, an anonymity tool of growing individual and societal importance.
- We uncover and describe important differences in how experts and non-experts understand and conceptualize Tor. Specifically, we show that gaps and inaccuracies in non-expert understanding of the operation and threat model of Tor could lead to a sense of more or less privacy and security than is actually the case.
- We suggest solutions that can improve the Tor user experience and boost adoption by non-experts, many of whom are in vulnerable situations and/or serve as society's important actors.

In the next section, we summarize prior research on the usability of Tor as well as that of privacy and security tools, in general. We then outline the method we used to conduct our study along with the details of participant recruitment and a description of the sample. Next, we describe our findings followed by a discussion of the insight that emerged. We proceed to apply the insight to suggest a number of potential improvements to Tor and other related aspects. We conclude after pointing out important limitations and avenues for future work.

2. RELATED WORK

There is a vast body of work on the technical aspects of Tor, such as attacks, defenses, case studies, etc. [10,11,27,33]. In contrast, our focus is on Tor users and their user experience. In this regard, we first present existing research that specifically targets the user experience of Tor, followed by a summary of the literature on user experience considerations in privacy and security tools, in general. We highlight the lack of research attention to studying Tor users and their motivations and practices.

2.1 User Experience of Tor

As Dingledine and Mathewson [3] observed, user-centered security [35] is important for anonymity systems since improving the user experience attracts more users, which strengthens the network as a whole. To this end, studies of the user experience of Tor have covered software and network operation, user interface, and external factors.

One of the first studies regarding the user experience of anonymity systems introduced latency 'shocks' into the anonymity network *'AN.ON'* over a one month period. A latency shock occurred every 105 minutes and lasted 15 minutes [15]. The results showed that the number of users who leave an anonymity network because of latency is linearly related to the amount of latency, for latency periods lasting less than 60 seconds. Fabian et al. [6] applied metrics from the literature to investigate and quantify such losses in usability caused by the latency within the Tor network. When compared with direct connections, they found that the median load time for a Web page over Tor was 5 times higher and Domain Name System (DNS) requests were 40 times slower. Based on these measurements, they postulated a request cancellation rate of 74%, leading to potential user frustration when using Tor. Given the negative impact of latency on the user experience, understanding and fixing the causes of latency within the Tor network is an important ongoing concern of Tor developers [5].

Other studies have examined the user experience of the various user interface elements of Tor. Clark et al. [1] performed a cognitive walkthrough of four configurations of the Tor software, performing four tasks in each of the configurations. They proposed user interface changes based on the difficulties encountered in completing the tasks. Norcie et al. [23] tried to identify the challenges experienced by individuals in adopting and using Tor, beginning with the step of installing the software. Their study of 25 undergraduates found that 64% of the participants faced various problems in installing and using the Tor Browser Bundle to perform the given tasks. These problems included difficulties finding and downloading the installation program, issues with decompressing the installation file, confusion in distinguishing between the Tor Browser Bundle and Firefox, latency, etc. In a follow-up study, Norcie et al. [22] evaluated the effectiveness of their proposed interface solutions aimed at fixing the problems uncovered in their initial study. They found statistically significant usability improvement in the case of most issues. Similarly, Lee et al. [17] examined the usability of the Tor Launcher that configures Tor connections. They found that the Tor Launcher interface required users to understand technical terms and did not provide appropriate and adequate feedback, thus leading to frustration and errors. They further showed that interface changes to the Tor Launcher were effective in addressing these challenges.

In a different vein, Khattak et al. [14] investigated how the Tor user experience is affected by the actions of external parties. Specifically, they looked at how Tor users are treated at the application as well as the network layer. They discovered that 1.3 million IPv4 addresses and 3.67% of the Alexa top 1,000 websites offered degraded services to Tor users or blocked them altogether.

2.2 User Experience of Privacy and Security Tools

At a more general level, researchers have devoted attention to the user experience of various commonly used privacy and security tools and mechanisms. We highlight the most salient findings in this domain pertaining to expert and non-

[4]https://securedrop.org

expert understanding and behaviors.

Leon et al. [18] studied 9 tools designed to limit or prevent online behavioral advertising and found significant usability problems in all of them, making it difficult, if not impossible, for users to make meaningful opt-out choices. Wash [31] and Wash and Rader [32] described variations in user mental models regarding viruses and hackers and explained that user decisions to follow security guidance from domain experts were influenced by the specifics of these mental models. Ion et al. [9] found that security non-experts deferred or ignored installing software updates, did not employ two-factor authentication, and did not use a password manager. They suggest that better messaging and usability are required to address the lack of adoption of common security tools. Similarly, Kang et al. [12] reported large differences in the complexity of the mental models of tech savvy participants and others. Yet, they found no link between technical knowledge and attempts to control online privacy. McGregor et al. [20] focused on journalists, a user group that often encounters situations that require anonymity, for sources as well as themselves. Journalists from the US and France indicated resorting to ad-hoc security approaches due to the lack of comprehensive and usable tools and reported difficulties in authenticating sources using existing tools.

2.3 Tor Users

While several of the studies mentioned above focused on the Tor *system*, very few of them attempted to understand Tor *users*. McCoy et al. [19] analyzed the traffic from an entry guard and an exit node under their control and found that a disproportionate number of users of the Tor network hailed from Germany, Turkey, and Italy. Additionally, they uncovered that notable amounts of sensitive information was sent as plain text over insecure protocols. In contrast to such indirect indicators of Tor user practices, we present accounts of Tor use obtained directly from the users themselves. Additionally, we discuss user motivations for adopting Tor and describe user understanding of Tor operation and threat model.

A recent survey reported that 34% of a sample of American adults who were aware of government surveillance programs took steps toward protecting their online information from the government [24]. Yet, only 2 of these people reported using anonymity software such as Tor, highlighting the huge gap between the expressed need for anonymity systems and their adoption in practice. Based on interviews of 17 current Tor users, we suggest user experience improvements that could help broaden its adoption.

3. METHOD

To address our research questions, we conducted semi-structured interviews with individuals who reported using Tor. The subsections below describe how we recruited participants and provide the details of our study protocol. The protocol was approved by New York University's Institutional Review Board (IRB).

3.1 Recruitment

Recruiting Tor users for such a study is difficult because only a small proportion of the population uses Tor. Moreover, Tor users are likely privacy conscious and, as a result, may be unwilling to discuss their attitudes and behaviors, especially pertaining to their use of Tor. Therefore, we cast a wide net and utilized multiple channels to seek study participants. Such an approach was also aimed at increasing the diversity of the sample. Specifically, we advertised the study on the Tor community of Reddit,[5] the 'Et Cetera Jobs' category of Craigslist for the New York City area, and mailing lists and bulletin boards at New York University. When describing the study on Reddit's Tor community and at the university, we mentioned that the research was regarding Tor. In contrast, on Craigslist, we stated that we were studying software use, without specifying our focus on Tor. This dual strategy was adopted partially to overcome the difficulties of attracting Reddit and university participants for a general software study and partially to include participants with varying levels of familiarity and experience with Tor. Our Craigslist advertisement directed potential participants to a brief online screening questionnaire (see Appendix A). Along with age, gender, and email address, the questionnaire asked about the use of 14 technologies and online services, with 'anonymization software' as one of the options in the randomly ordered list. Those who indicated using anonymization software were contacted to ask if they had ever used Tor.

3.2 Participants

We set up interviews with the individuals who reported having used Tor and expressed willingness to participate in the study. Overall, we interviewed 17 participants (5, 2, and 10 via Reddit, university channels, and Craigslist, respectively): 10 males, 5 females and 2 who preferred not to reveal their gender. Apart from ensuring that each participant was above the age of 18, we did not collect age information in order to respect the privacy and anonymity of the participants.[6] Participant occupations covered a spectrum of technical sophistication from penetration tester to fitness trainer. As a token of appreciation for participating in the study, we offered each participant a $20 gift card for Starbucks. Many participants declined the reward, likely to preserve their anonymity.

3.3 Study Protocol

Prior to participation, we provided the participants with information on the purpose of the study along with the procedures followed for handling the collected data. Specifically, we stated that we would not collect any personally identifiable information and would treat all responses as anonymous and confidential.

After obtaining informed consent for participation (and optionally for audio recording the conversation), we interviewed the participants one-on-one using a semi-structured interview protocol (see Appendix B). When possible, interviews with the participants local to the New York City area were conducted in person at New York University. Others were interviewed via phone or conferencing software, with the exception of one participant interviewed via email[7] and

[5] https://reddit.com/r/Tor

[6] Based on the responses to the screening questionnaire and our interactions with the participants, we estimate the age range to be 21–50.

[7] The questions were sent to the participant in an initial email, with subsequent emails used to ask follow-up questions as necessary.

Tasks	Internet Service Provider	Government and Law Enforcement	Target Web site or Service	Advertising Networks
Browsing a Web site	Can see one is using Tor	Can potentially see one is using Tor	Can see some Tor user is visiting the site	Can see some Tor user is visiting the site
Reading email	Can see one is using Tor	Can potentially see one is using Tor	Can access identity and data, but not IP	Can see some Tor user is visiting the site
Receiving an advertisement	Can see one is using Tor	Can potentially see one is using Tor	Can see some Tor user is visiting the site	Can see some Tor user is visiting the site

Table 1: An empty version of the above table was presented to the participants during the interview. Participants were instructed to fill out the cells indicating which information about them they believed the corresponding entities could access when they performed the listed tasks with the Tor Browser Bundle. The above table shows the correct answers derived from the Tor Project documentation [30].

two others interviewed using a text chat program.[8] The first author conducted all interviews.

Each interview consisted of several open-ended questions. At the beginning, the participants were asked general questions about their occupation to make them feel at ease and establish rapport. After the introductory questions, the interview delved into the participants' use of Tor, beginning with how they discovered Tor and covering the details of why, where, when, and how they used Tor. We further asked the participants to describe their understanding of how Tor works.

For an elicitation of the participants' understanding of the underlying operation of Tor, we asked them to engage in a drawing task as suggested by Kearney et al. [13]. Specifically, we asked the participants to draw a free-form sketch of their views and understanding of Tor, including its various front- and back-end (i.e., visible and invisible) components, processes, and actors. We stated that the sketches may include information about data flows and access controls. As they drew, the participants were encouraged to vocalize their thoughts in order to allow the collection and comprehension of the corresponding detail. Those who were interviewed via phone, conferencing, chat, or email were asked to send a picture of the drawing to the interviewer. When needed, we sought clarification and asked follow-up questions during the task. All drawings were retained for analysis.

We next asked the participants to fill out a table to capture their awareness of the threats countered by Tor (see Table 1). The table included a set of tasks along with various entities involved in those tasks. The participants were asked to indicate which pieces of information each of these entities could access when they used the Tor Browser Bundle to carry out each of the listed tasks. We encouraged the participants to think aloud when filling out the table. These answers, coupled with the responses to the other questions, allowed us to determine the participants' understandings of the potential deanonymization risks.

At the end, we asked the participants about the societal role of privacy tools, specifically in relation to contemporary national security debates and discussions in the US and Europe. We concluded the interviews with a brief multiple-choice questionnaire that used 5 questions on cybersecurity and anonymity taken from the 'Technical Knowledge of Privacy Tools Scale' from Kang et al. [12]. We chose this scale due to its topical relevance as well as short length. Participants who provided no more than one incorrect answer were marked as 'experts' with the remaining labeled 'non-experts.' These cutoffs were determined based on prior pilot testing with privacy and cybersecurity domain experts. Overall, 6 of our participants were classified as experts and the other 11 were treated as non-experts.

Most interviews lasted approximately 45 minutes. For the interviews that were audio recorded, the audio files were labeled with an anonymous identifier and destroyed after transcription. We analyzed the text of the interview responses along with the corresponding interviewer notes and the sketches collected during the drawing task. We followed an inductive process, allowing insight to emerge from the collected data. In order to avoid biasing the inductive analysis, we deferred a systematic review of the literature related to mental models of security and privacy tools until after the analysis was completed. The analysis included iterative open coding, axial coding, and selective coding [7] using the Atlas.ti software.

The first author began the three stages of coding – open, axial, and selective – right after the first interview. The coding proceeded continuously as the interviews were being conducted. During open coding, the text was coded sentence by sentence. Codes were created from the data with no initial hypotheses. For example, the sentence *"curiosity; I heard a lot of different things about it and was wondering how it works"* was labeled with the code 'being curious.' Axial coding examined the collection of codes generated by open coding and grouped related codes into categories. For instance, the codes 'feeling less watched,' 'feeling at ease,' 'evading surveillance,' and a few others were categorized under 'benefits derived from Tor use.' We further examined how frequently codes were mentioned together. Finally, in selective coding, the interactions between the categories and the codes were analyzed qualitatively and, to a smaller extent, quantitatively. The following sections describe the high level insight regarding user perceptions and understandings of Tor that emerged from the analysis.

4. FINDINGS

Unsurprisingly, we found notable and large differences between the experts and the non-experts in terms of understanding of the operation of Tor as well as the threat it counters. The experts exhibited deep knowledge of Tor's underlying operation while the views of the non-experts were simple and abstract. Notably, not all experts were free of

[8] These participants did not wish to reveal their voice and demanded a text communication channel with end-to-end encryption.

gaps in knowledge that could potentially affect anonymity during Tor use. Interestingly, the experts focused on the *technical details* of Tor operation, while the non-experts were much more likely to situate Tor within a broader *sociotechnical* landscape of purposes, actors, and values. We unpack these results by discussing the details of the participants' understanding of Tor operation and threat model, respectively.

4.1 Mental Models of Tor Operation

As mentioned above, we uncovered differences in the mental models of the experts and the non-experts pertaining to how Tor operates as a system. However, within each of the two participant groups, the models exhibited common threads.

4.1.1 Experts View Tor as a Complex Network

The experts understood Tor as a complex decentralized network used to move packets of information from one node to another. When describing how Tor works, the experts focused on network related aspects, such as connections, paths between Tor nodes, routing, etc., along with technical details, such as encryption layers. For example, one expert discussed the evolution in his understanding of Tor operation using the technical jargon of computer networks:

"When I started off I understood [Tor] pretty crudely as just kind of a way to get past state firewalls and to hide your identity from Web sites you are visiting. As I continued to use it, it's really good for NAT [Network Address Translation] traversal for example. Like, if you want to host a Web site from your home address and you're behind NAT, a Tor hidden service is a great way to give you that kind of access." (P8, Expert, Male)

Typically, the experts viewed the Tor network as composed of three elements: a sender, a receiver, and a path of decentralized nodes connecting the sender and the receiver. Moreover, they frequently referred to themselves as the sender who uses the network of Tor nodes to send messages to various receivers. For example, consider expert P10's sketches of Tor operation; he drew two diagrams, one depicting a connection from himself to a 'clearnet' site (see Figure 1) and another showing his connection to a Tor Onion Service (see Figure 2). In the first drawing, P10 indicated how relay information is loaded (including the possibility of a Tor bridge with obfuscation). The bottom half of the drawing shows that the traffic between the client (User) and the exit relay (Exit Node) is encrypted (green) and the traffic between the exit relay (Exit Node) and the Web site (Clearnet Site) is potentially unencrypted (red). In the second drawing, P10 showed the role of Tor Onion Service Directories, Rendezvous Points, and Introduction Points in connecting to a Tor Onion Service (Hidden Service). These drawings and descriptions present a mental model of the Tor network that demonstrates an understanding of the Tor system architecture akin to that of a Tor developer or researcher.

Other experts described Tor operation in varying levels of detail, with P10's being the most descriptive and complete. Despite differences in the level of completeness of the descriptions, all elicitations of the experts referred to the decentralized network nature of the Tor system architecture along with the role played by onion routing and encryption in the operation of Tor. For instance, the experts discussed the workings of Tor in terms of technical mechanisms, such

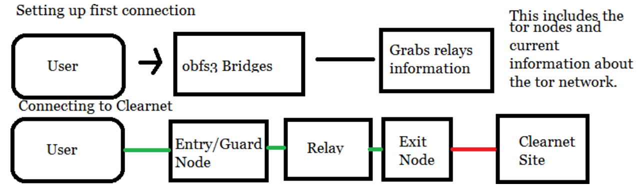


Figure 1: An expert's sketch of Tor's connection to a 'clearnet' Web site. (P10, Expert, Male)

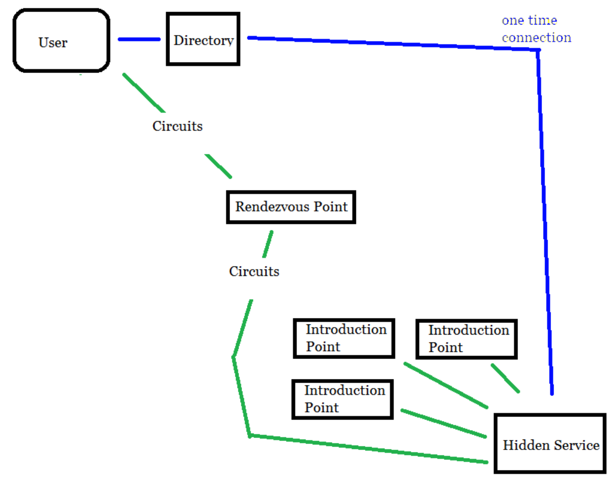


Figure 2: An expert's sketch of Tor's connection to a Tor Onion Service (Hidden Service). (P10, Expert, Male)

as traffic obfuscation techniques, anti-tracking measures, latency reduction solutions, etc.

4.1.2 Non-experts Treat Tor as a Service

Seven of our non-experts began using Tor out of curiosity. This curiosity took different forms, with four curious about the 'Deep Web' and controversial hidden services and others about the ability to surf anonymously or bypass censorship. Similar to the experts, the non-experts viewed themselves as information senders within the Tor system. However, unlike the experts, the non-experts often treated several key components of Tor's network based architecture as an abstract and opaque 'black box' with certain inputs and outputs. Specifically, we noted that the non-experts tended to treat Tor as a 'service.' They described calling upon the Tor service to perform specific functions, such as "*bouncing signals*" (P3, Non-expert, Male) or "*providing security*" (P11, Non-expert, Male). For instance, non-expert P17 drew his model of Tor as a service that provides a "*new me,*" obscuring his identity from those he is connecting to (see Figure 3). Additionally, Figure 3 reveals that the non-experts often mistakenly understood the Tor 'service' as *centralized*, with an administrator watching over and controlling the operation of individual Tor nodes. Only one non-expert correctly mentioned the decentralized nature of Tor nodes.

Different non-experts believed that the Tor service performed different functions; some said it provided security, others mentioned it made them anonymous, and still others stated it granted them access to previously inaccessible sites and resources. These functions were seen as enabling Tor to help the user achieve specific goals and tasks. These included tasks such as visiting sites that the participant wished to

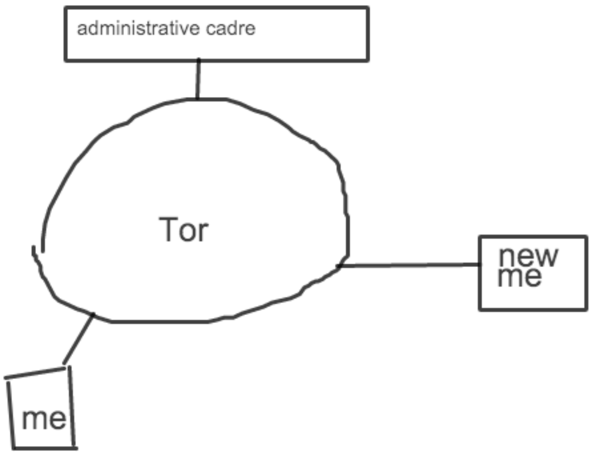


Figure 3: One non-expert's sketch describing Tor as a service with an administrative section watching over its inner workings. (P17, Non-expert, Male)

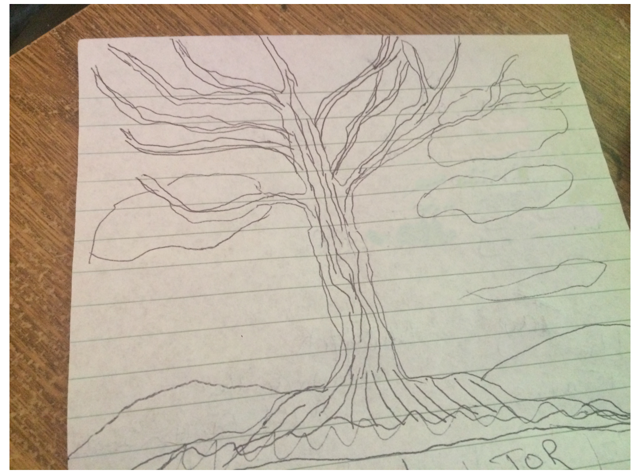


Figure 4: One non-expert's sketch depicting Tor as the Tree of Knowledge. (P9, Non-expert, Female)

conceal from the spouse, accessing geographically restricted content, circumventing content restrictions of filters and firewalls, etc.

While all non-experts described Tor as an abstract service, some descriptions exhibited more technical sophistication than others. For example, one participant mentioned that Tor may assign a new IP address, showing some understanding of the role of an IP address as an identifier.

"And then like IP address . . . I don't know . . . does Tor jumble up your IP? Maybe, perhaps it does, perhaps it doesn't. Perhaps it gives you a new IP." (P2, Non-expert, Female)

Three non-experts mentioned cryptography, even though they did not understand the role it played in the operation of Tor. Two non-experts mentioned 'signal bouncing' without explaining how it was accomplished.

"Like, the signal gets split up among other things, that would be cool if that happens, not too sure how that works, but I don't have an extensive knowledge of that." (P2, Non-expert, Female)

"As far as I'm aware the way it works is it bounces your signal around a lot . . . To various countries and such." (P3, Non-expert, Male)

It should be noted that there was a large degree of uncertainty among the non-experts about their understanding of the operation of Tor. While some non-experts were confident in their answers, five seemed unsure that their understanding was accurate or complete. For instance, when P2 was asked to clarify her idea that Tor performs *"signal dispersion,"* she replied that it works with *"cryptography,"* admitting that she did not know what that meant, indicating confusion between terminology and operation. Other non-experts simply stated that they did not understand how Tor worked, but knew that it did.

"It's one of those things where I know it works, it exists." (P9, Non-expert, Female)

Five non-experts described their understanding of Tor operation through metaphors. For example, one non-expert stated that it worked just like a faucet: *"if one turns the handle, the water appears"* (P9, Non-expert, Female). P11 clarified his sketch of Tor by equating it with Fort Knox, through which his traffic passed in order to become secure. This demonstrates his conception of Tor as a central service meant to secure, rather than anonymize, his traffic.

"Let's say I am like a circle. I am a circle on the left side. Inside of the circle I have for example, let's say I have my laptop and this for example would be down, hanging down. And on this it says it's my computer or it's my laptop. So that's on the left side and above that for example you can put any human picture and I give it a face. In the middle for example you will have a wall like Fort Knox and that would be in the middle obviously with no face because . . . it's not human and on the right side it is also a circle and that would be another computer with another human face." (P11, Non-expert, Male)

Many metaphors utilized by the non-experts described the ideologies and the values that the participants believed Tor stands for. For instance, P14 referred to Tor as the Statue of Liberty, bringing liberty to those who use it.

"Yes, the Statue of Liberty on Ellis Island. So just to describe to you we can probably explain it as giving us liberty to watch what I need, you know, and so at the same time [providing] freedom." (P14, Non-expert, Male)

Another example is a non-expert drawing Tor as the Tree of Knowledge (Figure 4), granting access to many different branches of knowledge.

"Tor ties in with the Tree of Knowledge for the simple reason that it's one of the best forms of confronting knowledge. Because there's no filters really on Tor." (P9, Non-expert, Female)

Apart from underscoring the non-expert treatment of Tor as a service, these metaphors also reveal that the non-experts often viewed Tor as a tool for social good. This aspect was mentioned in multiple non-expert interviews, with the participants discussing Tor as a tool used by activists, journalists, and ordinary citizens for communicating freely without surveillance, bypassing state censorship, and achieving empowerment in civic engagement.

4.2 Threat Model Addressed by Tor

During the interviews, we attempted to discover the participants' understandings of the threat model of Tor. We discovered misunderstandings of the following threats to anonymity on Tor:

1. **Client side scripting:** Client side scripting may place users at risk. For example, Flash code running outside the browser's control can be used to deanonymize users. Similarly, various vulnerabilities in JavaScript running within the browser can be exploited for deanonymization.
2. **Browser fingerprinting:** When Tor users use a browser other than the Tor Browser Bundle over the Tor network, the browser sends information to visited sites, such as installed add-ons, version, etc. Since the number of people with matching sets of information is likely to be low, the browser fingerprint lowers anonymity, with the worst case being unique identification.
3. **Side channel leaks:** Information provided by users to third parties external to Tor, such as login credentials, credit card numbers, or even language choice, can be used to denanonymize the user to varying degrees. In addition, if users do not ensure the use of encrypted connections, their information can be accessed by the exit node on their Tor circuit.
4. **Node operation:** Tor nodes are independently owned by volunteers. As a result, data flowing within the network is not controlled or seen by any single party, including the owners and the operators of the Tor Project itself. As a result, it is possible for malicious actors to run Tor nodes with the aim of attacking users who utilize the node (which is typically an exit node).

Similar to the operational details of the Tor system, the experts and the non-experts differed in the understanding of the threat model that Tor addresses. We discuss each in turn.

4.2.1 Experts Mostly Grasp the Threat Model

The experts showed a reasonably accurate understanding of the threat that Tor attempts to counter. Importantly, they understood that Tor is not a *complete* solution for all potential anonymity related issues and additional steps may be needed to achieve the desired level of anonymity. For instance, when filling out Table 1, expert responses revealed that they understood the complexities of the different browsing tasks and situations. These complexities are tied to the threat model of Tor. For example, all experts understood that logging into a Web site could deanonymize them. When asked whether the email service could access any information when reading email using the Tor Browser Bundle, the response of one expert demonstrated his understanding of the limits of Tor's protection:

"Yes they do, because you have an account with them. Assuming you've provided personal information, they kind of know who you are and, you know, what you've sent, but they still don't know where you are. You've still obscured your IP address." (P1, Expert, Male)

Further, all experts mentioned that the traffic exiting a Tor exit node may not be encrypted, again demonstrating the limitations of the protection Tor provides.

"Between the laptop and the entry I will write a little note that says ISP can see that I'm using Tor. And then between the entry and the middle I'm going to say 'encrypted traffic.' And then between the middle and the exit I'm going to say 'encrypted traffic' and between the exit and the Web site I'm going to say 'ISP can see requests, but not the originator.'"' (P1, Expert, Male)

In addition, many experts understood that the threat model of Tor allows a certain number of compromised Tor nodes, and that some of the Tor nodes might be a threat.

"Nodes may be owned/0wned[9] by governments." (P5, Expert, Unspecified gender)

Though the experts understood the threat model, all but two of them neglected to mention the Tor Browser Bundle as a part of the Tor system, mentioning only its network elements, such as the nodes, and security elements, such as encryption. Two experts configured their own Web browsers or used other non-standard ways to connect to the Tor network to receive Web content. This makes them vulnerable to fingerprinting attacks mentioned above, thus leading to potential deanonymization. Moreover, using a Web browser other than the Tor Browser Bundle is complicated and could lead to mistakes such as DNS leaks caused by a misconfigured browser resolving DNS requests independent of Tor. One expert stated that he used *wget* (an alternative tool for Web content retrieval) over Tor, which has a similar effect if the user does not anonymize the USER-AGENT string.[10]

4.2.2 Non-experts Conflate Threat Models

Unlike the experts, the responses of the non-experts revealed a lack of consensus regarding the threats that Tor addresses. While some non-experts possessed a complete understandings of the Tor threat model, five believed that Tor provided more security than it actually does. For instance, one non-expert believed that Tor was a tool for protecting sensitive data, such as credit card numbers, in transit on the Internet.

"It's going to something and entering my credit card or some kind of financial or some Web site where I don't want them to have my information because they're going to follow me." (P13, Non-expert, Female)

Another non-expert believed that Tor kept one anonymous from one's email provider, even when logged into the service. Other non-experts, however, held the view that Tor did not offer complete protection, with four claiming that Tor is effective for privacy protection from entities such as advertising networks, but not from governments and ISPs. Two others believed that the Tor Project has access to all traffic on the Tor network and could provide it to governments and law enforcement agencies. One non-expert argued that the Tor Project does not provide such access only because doing so would be counter to their goals.

"I know that I'm not doing anything dangerous but they don't know that, so I can see why the government would want to

[9]0wned here refers to the computing slang term indicating a device being taken over and controlled by an external party, with or without the knowledge of the device owner.

[10]A USER-AGENT string is a line of text containing information about the browser or the program.

have access to that kind of thing. Or maybe they can receive alerts from Tor saying 'hey this person is suspicious by your standards' ... but that's bad business, so..." (P4, Non-expert, Female)

Two non-experts claimed using Tor to circumvent geographical restrictions imposed by Web sites, such as Hulu, Netflix, etc. Yet, many of these sites run Adobe Flash or JavaScript, which can not only deanonymize users but also leave them vulnerable to injection attacks from malicious Tor exit nodes.

In general, the non-experts operated with incomplete, and sometimes inaccurate, understanding of the Tor threat model, often conflating it with other threat models that Tor is not designed to address. These gaps and inaccuracies could lead to a sense of more or less anonymity and privacy than is actually the case.

4.3 Discovery and Use of Tor

We examined how the participants discovered Tor, why they used it, and how long they had been using it. There are no real distinctions between the experts and the non-experts regarding the discovery of Tor. Both groups primarily discovered Tor through news articles, and many participants reported discovering it around the time of the initial publication of the Snowden documents. Some exceptions exist, with five participants finding Tor through searches on popular search engines or hearing about it from friends. One participant discovered Tor at a conference, and two participants (both experts) did not remember how they discovered Tor. In terms of use, however, we found significant differences between the experts and the non-experts.

4.3.1 Experts Used Tor for Many Reasons

All experts reported that they used Tor more frequently and for more purposes than the non-experts. A few experts used the Tor Browser Bundle as their primary browser, using it for most tasks and reserving non-anonymous browsers, such as Google Chrome and Mozilla Firefox, only for tasks which are ill-suited for the latency Tor creates (e.g., video streaming, etc.).

"I use [Tor] primarily as my everyday browser for most of my tasks. But I use regular Firefox if I want to do something, if the Web site is blocking Tor or if I want to do something on localhost that doesn't need outside Internet access." (P1, Expert, Male)

In addition to anonymous browsing and censorship circumvention, the experts mentioned alternative uses of Tor apart from Web browsing, such as downloading via alternative means such as *wget*, circumventing NAT using Onion Services, etc.

"So if I'm at school I can use Tor to ssh into a computer on my home network and it's not a problem. I don't have to deal with all of the IP address stuff." (P8, Expert, Male)

Curiosity differed between the experts and the non-experts. The experts tended to be curious about the network and its components, rather than the information held in Onion Services.

"Pure curiosity drove me toward it. It was just a different way of distributing information systems, so it was like, hmm, if we could do it a bit differently that would be a bit better." (P10, Expert, Male)

Additionally, the experts who started using Tor out of curiosity tended to remain Tor users and become more involved in the Tor community, while the non-experts who started using Tor due to curiosity stopped using it relatively quickly.

4.3.2 Non-experts Have Specific Motivations

Although the non-experts mentioned a variety of reasons for using Tor, all but two used it only within the context of a single specific purpose. Non-expert motivations for using Tor included: satisfying curiosity regarding the content accessible via Tor, bypassing censorship, circumventing geographical restrictions imposed by Digital Rights Management (DRM), countering surveillance by governments as well as other parties such as advertisers, communicating with activists, protecting the discovery of one's visits to pornography and gambling sites, researching sensitive legal matters, etc.

The non-experts who used Tor out of curiosity tended to be more curious about the information available via Tor rather than about the operation of the anonymity system itself. Specifically, the non-experts were drawn to information available on Onion Services, also called the 'Deep Web.'

"To be honest, the Internet black market. Uh, yeah, just to access it and see what's up. Um, the 'Deep Web.' Yes, that's it, the 'Deep Web.'" (P2, Non-expert, Female)

Four non-experts believed that Tor was designed primarily in the context of their own specific use case. For example, one non-expert used Tor only when abroad in a country that censored Web sites.

"I was using [Tor] because I was living abroad and I wasn't allowed to access certain sites... I was looking for ways to access these sites or rather looking for ways to get around the countrywide ban." (P4, Non-expert, Female)

She stated that Tor was not very needed in the US because the US government did not block many Web sites.

"I feel like it's less relevant in the US for the average user because the US doesn't block too much. They don't block Facebook and they don't block Google or that sort of thing. Whereas within a lot of foreign countries there's a lot of content that the US would consider benign that the governments wouldn't want you to access." (P4, Non-expert, Female)

Another non-expert used it only when performing credit card transactions, believing Tor to be a tool meant to safeguard data in transit.

Similarly, the non-experts tended to focus on only one adversary while using Tor. For example, one non-expert claimed that she used Tor because she did not want the government to see that she had looked up drugs, contract killer postings, and other such information.

"I just kind of used it those few times to look on the Internet and be like 'look how much acid costs on the Internet' and then like...find all the Web sites that are like oh I'm a hitman and I'm going to kill the president for a few million dollars... I like the president, but ... I was kind of like just lurking and seeing what's up. That was the main pur-

pose and I didn't really want to get like a knock on my door, which they do in China... So that's like something that I'd like to avoid, which I'm sure doesn't happen as frequently in the States but... I don't like want to get arrested for some unrelated incident and then have my record like.... my computer searched, and then its like 'You were looking at hitmen, what's up with that?"' (P2, Non-expert, Female)

Lastly, half of the non-experts reported using Tor infrequently or having quit using it altogether, citing a lack of need or fading curiosity for the tool as their reasons.

4.4 National Security and Tor

In contrast to other aspects, we found no major differences among the experts and the non-experts regarding the relationship between Tor and national security concerns. When asked about the morality of Tor and its role in national security, most participants stated that Tor was a trade-off between privacy and national security and acknowledged that it likely made law enforcement more difficult. Yet, all but one participant believed that Tor was a good tool and the balance between individual privacy and national security should be closer to privacy.

"On balance I think that the good parts outweigh the bad parts and that they are necessary regardless of what we might think of the bad parts. So obviously properly implemented secure communication technologies will always be problems for law enforcement and intelligence agencies because they depend on sort of exclusive access to our data as part of their job. But ... I mean that's fine but there are other things at stake, right? There's individual liberty, there's freedom of speech, there's freedom of association, there's the ability to have secure technologies that will protect really important sensitive information, you know embarrassing stuff or your credit card number." (P8, Expert, Male)

There were some exceptions, however. One participant believed that privacy and national security are synergistic, and the protection of the rights of the people, including privacy, is itself a matter of national security.

"In my opinion, security is directly related to privacy and so is privacy to anonymity. I feel stronger tools are needed and are a benefit to society. Giving up any of the three (security, privacy, anonymity) means you can have none of the above. I understand the national security threat when the 'bad guys' use these tools, but they won't follow the rules anyway." (P6, Expert, Unspecified)

Another participant believed that Tor was detrimental to national security and should include a back door that allows access to the government.

"For national security reasons there is a need to have back hole [back door] access to certain things ... Tor is something that can be a very positive tool but at the same time it is used by a lot of illegal entities ... everything from child pornography to black market smuggling to terrorism, finances, planning, and coordination and so in that sense I think that there needs to be a certain degree of control from a government perspective." (P17, Non-expert, Male)

It must be noted that the views of the non-experts on this matter may have been influenced by some of the misunderstandings described in Sections 4.1 and 4.2. Specifically, a few non-experts believed that intelligence agencies, such as the Central Intelligence Agency (CIA) and the National Security Agency (NSA), are capable of defeating the protection Tor provides and have access to Tor network traffic. As mentioned earlier, one participant believed that Tor was capable of giving notices to the government if a Tor user is deemed suspicious by government standards, but would not do so because of the business implications of such an action. Yet, most non-experts believed that Tor helped foster important sociotechnical values, such as freedom of speech, uncensored information access, privacy, and personal security.

Category	Expert	Non-Expert
Mental Model	Complex network	On-demand service
Threat Model	Multiple threats	Specific (single) threat
Frequency of Use	Frequent	Mostly for specific uses
Discovery	Varied	Mostly through news
Morality of Tor	Good, positive	Varied, mostly positive

Table 2: Comparison of notable aspects of the understanding and the use of Tor across the experts and the non-experts.

5. DISCUSSION

Table 2 summarizes the notable aspects of our findings across the experts and the non-experts. In addition to the findings related to our research questions, we found that several experts and non-experts mentioned enhancing their anonymity and privacy by engaging in 'compartmentalization' via the use of a separate device for Tor use. Such a practice indicates greater attention to privacy and security among Tor users in comparison with non-users. While the adoption of Tor in the general population remains low, our sample shows that its user base is heterogenous and not composed only of domain experts with deep technical knowledge.

As expected, our findings confirm that the extent to which non-experts grasp the operational details of Tor differs substantially from the level of understanding of experts. Non-expert understanding of the operational details of Tor varied widely, possibly because of the differences in the frequencies and the motivations of use. Our findings shed light on the nature of these differences in terms of mental models and threat models. Regardless of the technical sophistication of these mental models, Tor, like any privacy enhancing technology, would benefit greatly from understanding and utilizing the mental models of its users [31]. For instance, the user interface as well as the documentation of Tor could draw upon the mental models to present the operational concepts more effectively.

The experts exhibited useful and complete knowledge of the Tor architecture and operation along with a nuanced understanding of its threat model. In contrast, the mental models of the non-experts were incomplete and overly abstract, leaving out or distorting important details that impact anonymity and privacy. For instance, bounding the entire Tor network within a single box may create a false sense of privacy and security by ignoring the potential attacks by malicious exit nodes, such as capturing sensitive information passing through the node via insecure protocols [19]. Moreover, users operating under an assumption of anonymity may engage in behavior they might not want tied back to their identity. In contrast, viewing Tor as a centralized service could lead to the opposite effect. A belief that external parties, such as governments, law enforcement agencies, ISPs, and the Tor Project, can access decrypted Tor traffic has the potential to create a chilling effect, leading to self-censorship as well as unwillingness to use Tor. As mentioned earlier, even some experts exhibited gaps in understanding and engaged in behaviors that left them vulnerable to specific attacks, such as DNS leaks. This underscores that even the smallest of gaps in knowledge has the potential to defeat the anonymity protection a user seeks via Tor. Some of these issues, such as DNS leaks, can be addressed by the Tor software itself,[11] while others can be addressed by explicitly documenting the dangers of non-standard uses of Tor.

Tor is used by experts and non-experts across the world for a variety of purposes. Many of these purposes involve society's important values and causes, such as circumventing censorship, avoiding surveillance, sharing sensitive information of journalistic importance, communicating with informants, and so on. In addition, Tor serves sensitive and valuable personal purposes, such as protecting one's online activities from an abusive partner, avoiding targeted advertising, etc. In a large majority of these situations, the users involved are non-experts. In such circumstances, gaps and inaccuracies in the understanding of the operation and threat model of Tor that lead to deanonymization may hold serious repercussions, including account compromises, identity theft, financial losses (resulting from fraud), surveillance of communication and movements, civil or legal penalties, physical and/or psychological abuse, imprisonment, or, in extreme cases, death.

Interestingly, the responses of our non-experts show that they placed importance on the societal values that Tor aims to promote along with the corresponding usage scenarios tied to those values. Indeed, some of them seemed to be using Tor to make a value statement related to civil liberties and democratic principles, such as privacy, anonymity, freedom from surveillance, personal liberty, censorship circumvention, freedom of expression, etc. While the experts also recognized the connection of Tor to societal values, they preferred to describe Tor in terms of the architectural and engineering details of the software and the network. When considering whether Tor poses a problem for national security, participant opinions ranged from asserting that Tor acts as a force for freedom to believing that Tor is a tool for cybercriminals and terrorists.

Regardless of how they discovered Tor, the experts reported using Tor more frequently and for longer periods. In contrast, the interest of the non-experts tended to fade, with many claiming that they saw no need for the tool. While these usage differences have previously been observed in other privacy tools as well [25], they are especially crucial for an anonymity system such as Tor because the efficacy of its protection improves with an increase in the number of users. Given the awareness of Tor's value proposition exhibited by the non-experts, emphasizing that the use of Tor is a community and societal contribution could potentially boost its adoption.

A typical goal of Human Computer Interaction research is creating user experiences that facilitate effective use of a system without requiring deep knowledge of the underlying operation, thus making it easily accessible to non-experts. As discussed in Section 6, our findings can be applied to improve the Tor user experience for non-experts. However, a key aspect where Tor differs from typical systems is its use as a privacy and security tool, sometimes under circumstances of great importance as well as danger. As such, an incomplete or inaccurate understanding of its operational details has the potential for individual as well as societal harm. These risks lead to a tension between the need to promote technical understanding of the operational detail and the goal of making such knowledge unnecessary as a requirement for the correct use of the system. Addressing the issues uncovered by our findings could be a step in the direction of mitigating the potential risks and resolving the tension between the simultaneous needs for revealing as well as abstracting away the technical details of Tor operation.

6. IMPLICATIONS

Our findings can be applied to improve the Tor system in a variety of ways. These include refining the design of the user interface and the user experience of the Tor Browser Bundle, targeting specific operational aspects for enhancement and optimization, and facilitating learning, especially for non-experts. We discuss some of these below. In addition to these improvements, our data suggests that Tor users desire a reduction in latency.

6.1 Route Information

As discussed earlier, non-experts conceptualize Tor as a centralized service. A possible solution to avoid such a misunderstanding could be displaying information about the ownership of each Tor node in the current connection's route, when such information is available and verifiable. Such a feature would be an extension of the current Tor Browser Bundle functionality that allows clicking the onion logo to display route information, such as the node IP address and the country. It might also be useful to make such information readily available in the background without the need for explicit click-and-seek. The feature could be further expanded to indicate the encryption status of each link within the current route.

6.2 Safe Script Execution

Our findings suggest that a notable barrier to the adoption and the use of Tor is the demand and the need for using Web sites and services that utilize JavaScript. JavaScript is so ubiquitous that disabling it makes a large proportion of popular Web sites unusable [28]. As mentioned earlier, enabling JavaScript while using Tor may lead to deanonymization [29]. We advocate investigations of operational and architectural modifications that reduce the attack surfaces opened up by enabling JavaScript within the Tor Browser Bundle. Such technical improvements could facilitate a reasonable balance between preserving anonymity without

[11] For instance, DNS leaks can be addressed by raising a warning when the Tor network proxy receives a numeric IP address instead of a request for resolving a text based domain name.

overly compromising usability and utility.

In addition, the Tor Browser Bundle should be modified to warn users that it defaults to a *low* security level that has JavaScript enabled. In the current release, one must open a menu accessible by two clicks in order to discover this default setting. A prominent visual indicator of the current security level should be available at a glance in the Tor Browser Bundle interface.

6.3 Tor Friendly Web sites

In line with the empirical findings of Khattak et al. [14], most of our participants mentioned routinely having trouble due to the restrictions many Web sites place on Tor traffic. These sites typically restrict Tor traffic even under situations that pose minimal risk to the server, such as fetching static Web pages that do not involve user interaction or data input. We recommend that Web sites, especially those providing important information such as government Web sites, provide 'Tor friendly' versions of the pages that allow Tor users to at least fetch information, even if specific mechanisms, such as posting, are disallowed to protect against abuse. In addition, using CAPTCHAs to prevent abuse by Tor users should be limited only to submitting POST data, permitting GET requests without submitted parameters to proceed without such checks.

Moreover, site owners could consider serving an alternative Tor friendly version of the site for connections from Tor exit nodes. The process of creating such a Tor friendly version of a site could be made easier by promoting the creation of plugins for common Web development platforms, such as WordPress and Dreamweaver. Such plugins could ensure that Tor nodes are not blacklisted and automatically create versions of Web pages that reduce the amount of JavaScript to the bare minimum, or possibly none.

6.4 Compartmentalization

Our participants reported using the privacy enhancing strategy of compartmentalization by separating different tasks or personas through the use of separate computers and/or software. Yet, most current programs and operating systems make it challenging, if not impossible, to achieve meaningful compartmentalization of digital activities. We advocate explicit attention by system designers and Tor developers to the provision of compartmentalization functionalities as a privacy and security enhancing feature.

6.5 Maintaining Workflow

Although many participants in our study compartmentalized their Tor use, some participants indicated frustration at the burden of switching away from Tor in order to complete the tasks that could not be performed via Tor. These tasks included visiting Web sites that depend on flash or Java plugins or those that explicitly block Tor traffic. Currently, when a user wants to perform such tasks he or she must manually switch to another browser and copy/paste the site address. Subsequently, the user must remember to switch back to Tor once the task in the other browser is completed. While the task is ongoing, the user must switch back and forth between Tor and the other browser. The desire to minimize the disruption caused by the burden and frustration of managing the workflow and task switches can lead users to choose a non-Tor browser as the default. We suggest adding functionality within the Tor Browser Bundle that makes such task switches easier and faster whenever a task necessitates the use of a non-Tor browser. Such functionality has the potential to increase Tor adoption and usage by making it easier for users to stay within the Tor system as much as possible, switching away from Tor only when absolutely necessary for the task at hand. In addition to benefiting the individual user, increasing the time users spend using Tor would boost the overall utility of Tor by increasing the number of active users at any given time.

6.6 Contextual and Personalized Training

It would be beneficial to explore training and learning opportunities for non-experts in order to promote the development of useful conceptualization of the operation and threat model of Tor. Training has been shown to be effective in other cybersecurity domains, such as phishing [16]. In addition to an explicit focus on non-experts, training mechanisms could be customized to the person(s) and situation(s) at hand. For instance, different training modules could be developed for common use cases, such as circumventing censorship, avoiding surveillance, communicating securely and anonymously with a journalist, etc. Training activities could even be embedded into the user experience of the Tor Browser Bundle in a manner that utilizes learning theories and techniques, such as gradual knowledge building, periodic repetition, and effective assessment.

7. LIMITATIONS

A few limitations must be kept in mind when considering the generalizability of these findings. While we continued iterative coding of the interview responses until sufficient understanding emerged, we were unable to engage in purposeful additional sampling aimed at filling gaps. Although such a step is common in inductive qualitative analysis, the difficulties in finding and recruiting unbiased and unprimed Tor users limited our sampling efforts. Despite this limitation, we believe we reached reasonable saturation for the research questions at hand. In addition, the inherent difficulty in recruiting Tor users without bias or priming means that our study has a small sample size compared to other research on mental models. Further, advertising in Tor-specific groups such as Reddit's Tor community may have introduced bias in the sample. For privacy and anonymity, we did not collect demographic data beyond gender. As a result, we cannot account for cultural differences. Although we cannot be certain, advertising in online and offline communities in English leads us to believe that most of our participants were native residents of the US or Canada. Finally, we point out that our findings are derived from self-reports. Consequently, it is possible that the participants omitted, forgot, or misrepresented their understanding and behavior.

8. FUTURE WORK

To the best of our knowledge, we are the first to attempt to understand the mental models of users of any anonymity software. Our findings point to several opportunities for future sociotechnical research. In Section 6, we proposed potential solutions that involve changes to the Tor user interface and user experience. The effectiveness of these suggestions needs to be validated via empirical studies. In addition, we call for further design exploration in creating user experiences that balance the tension between revealing and abstracting the operational detail. Due to the qualitative nature of our study, the findings are derived from a small sample. To validate generalizability, an online questionnaire could be formulated based on these findings and administered to a larger sample covering a broader population. We focused our investigation only on the Tor anonymity system. Further research is needed to examine whether these findings apply to other anonymity systems, such as the Invisible Internet Project (I2P) [34] and Freenet [2].

9. CONCLUSION

Anonymity systems, such as Tor, are an important tool for providing privacy and security in a landscape of growing online surveillance and censorship. In addition to enabling ordinary citizens to assert their civil liberties, Tor serves as a crucial anonymity and safety mechanism for society's important actors, such as journalists, political dissidents, whistle blowers, human rights activists, etc. A large majority of these actors are not technical domain experts. We found that non-experts conceptualize Tor via abstractions and metaphors that hide important operational aspects, thus potentially compromising the anonymity they seek. In contrast, experts understand the underlying technical operation and threat model and are highly likely to possess an accurate understanding of the level of privacy and security protection afforded by Tor. Fostering useful and complete understanding of the operation and threat model of Tor is a critical need to avoid deanonymizing vulnerable users as well as to promote adoption of Tor.

10. ACKNOWLEDGMENTS

We thank the participants of the study. We are grateful to Lesley Fosh, Jeffrey Ramdass, Martin Shelton, and Rick Wash for insightful feedback on draft versions of the paper. We acknowledge the anonymous reviewers for helpful comments. Thanks are due to Thomas Davis and Mihir Mahajan for editorial input. This work is supported by National Science Foundation (NSF) grant DGE-0966187.

11. REFERENCES

[1] J. Clark, P. C. van Oorschot, and C. Adams. Usability of Anonymous Web Browsing: An Examination of Tor Interfaces and Deployability. In *Proceedings of the 3rd Symposium on Usable Privacy and Security (SOUPS 2007)*, pages 41–51. ACM, 2007.

[2] I. Clarke, O. Sandberg, B. Wiley, and T. W. Hong. Freenet: A Distributed Anonymous Information Storage and Retrieval System. In *Proceedings of Designing Privacy Enhancing Technologies: International Workshop on Design Issues in Anonymity and Unobservability*, pages 46–66. Springer, 2001.

[3] R. Dingledine and N. Mathewson. Anonymity Loves Company: Usability and the Network Effect. In *Proceedings of Workshop on the Economics of Information Security (WEIS 2006)*, pages 547 – 559. Springer, 2006.

[4] R. Dingledine, N. Mathewson, and P. Syverson. Tor: The Second-generation Onion Router. In *Proceedings of the 13th Conference on USENIX Security Symposium (USENIX Security 2004)*, pages 21–21. USENIX Association, 2004.

[5] R. Dingledine and S. J. Murdoch. Performance Improvements on Tor or, Why Tor is slow and what we're going to do about it. `https://www.torproject.org/press/presskit/2009-03-11-performance.pdf`. Accessed: 2017-06-15.

[6] B. Fabian, F. Goertz, S. Kunz, S. Müller, and M. Nitzsche. Privately Waiting – A Usability Analysis of the Tor Anonymity Network. In *Sustainable e-Business Management: Proceedings of the 16th Americas Conference on Information Systems (AMCIS 2010)*, pages 63–75. Springer, 2010.

[7] B. G. Glaser and A. L. Strauss. *The discovery of grounded theory: Strategies for qualitative research.* Transaction Publishers, 2009.

[8] D. M. Goldschlag, M. G. Reed, and P. F. Syverson. Hiding routing information. In *Proceedings of the International Workshop on Information Hiding*, pages 137–150. Springer, 1996.

[9] I. Ion, R. Reeder, and S. Consolvo. "... no one can hack my mind": Comparing Expert and Non-Expert Security Practices. In *Proceedings of Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 327–346. USENIX Association, 2015.

[10] A. D. Jaggard, A. Johnson, S. Cortes, P. Syverson, and J. Feigenbaum. 20,000 in league under the sea: Anonymous communication, trust, MLATs, and undersea cables. In *Proceedings on Privacy Enhancing Technologies (PoPETS 2015)*, pages 4–24. De Gruyter Open, 2015.

[11] R. Jansen, F. Tschorsch, A. Johnson, and B. Scheuermann. The sniper attack: Anonymously deanonymizing and disabling the Tor network. In *Proceedings of the Network and Distributed System Security Symposium 2014 (NDSS 2014)*. Internet Society, 2014.

[12] R. Kang, L. Dabbish, N. Fruchter, and S. Kiesler. "My Data Just Goes Everywhere": User Mental Models of the Internet and Implications for Privacy and Security. In *Proceedings of the Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 39–52. USENIX Association, 2015.

[13] A. R. Kearney and S. Kaplan. Toward a methodology for the measurement of knowledge structures of ordinary people: The conceptual content cognitive map (3CM). *Environment and Behavior*, 29(5):579–617, 1997.

[14] S. Khattak, D. Fifield, S. Afroz, M. Javed, S. Sundaresan, V. Paxson, S. J. Murdoch, and D. McCoy. Do You See What I See? Differential Treatment of Anonymous Users. In *Proceedings of the Network and Distributed System Security Symposium 2016 (NDSS 2016)*. Internet Society, 2016.

[15] S. Köpsell. Low Latency Anonymous Communication – How Long Are Users Willing to Wait? In *Proceedings of Emerging Trends in Information and Communication Security (ETRICS 2006)*, pages 221–237. Springer, 2006.

[16] P. Kumaraguru, J. Cranshaw, A. Acquisti, L. Cranor, J. Hong, M. A. Blair, and T. Pham. School of phish: A real-world evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security (SOUPS 2009)*, pages 3:1–3:12. ACM, 2009.

[17] L. Lee, D. Fifield, N. Malkin, G. Iyer, S. Egelman, and D. Wagner. A Usability Evaluation of Tor Launcher.

In *Proceedings on Privacy Enhancing Technologies (PoPETs 2017)*, pages 87–106. De Gruyter Open, 2017.

[18] P. Leon, B. Ur, R. Shay, Y. Wang, R. Balebako, and L. 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 2012)*, pages 589–598. ACM, 2012.

[19] D. McCoy, K. Bauer, D. Grunwald, T. Kohno, and D. Sicker. Shining Light in Dark Places: Understanding the Tor Network. In *Proceedings of the 8th International Privacy Enhancing Technologies Syposium (PETS 2008)*, pages 63–76. Springer, 2008.

[20] S. E. McGregor, P. Charters, T. Holliday, and F. Roesner. Investigating the Computer Security Practices and Needs of Journalists. In *Proceedings of the 24th USENIX Security Symposium (USENIX Security 2015)*, pages 399–414. USENIX Association, 2015.

[21] H. Nissenbaum. The meaning of anonymity in an information age. *The Information Society*, 15(2):141–144, 1999.

[22] G. Norcie, J. Blythe, K. Caine, and L. J. Camp. Why Johnny Can't Blow the Whistle: Identifying and Reducing Usability Issues in Anonymity Systems. In *Proceedings of the 2014 Workshop on Usable Security (USEC 2014)*. Internet Society, 2014.

[23] G. Norcie, K. Caine, and L. J. Camp. Eliminating Stop-Points in the Installation and Use of Anonymity Systems: A Usability Evaluation of the Tor Browser Bundle. In *5th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETS 2012)*. PETS Symposium, 2012.

[24] L. Rainie, M. Shelton, and M. Madden. Americans' privacy strategies post-Snowden. *Pew Research Center*, March 2015.

[25] K. Renaud, M. Volkamer, and A. Renkema-Padmos. Why Doesn't Jane Protect Her Privacy? In *Proceedings of Privacy Enhancing Technologies (PoPETs 2014)*, pages 244–262. Springer, 2014.

[26] B. Schneier. How to Remain Secure Against the NSA. `https://www.schneier.com/blog/archives/2013/09/how_to_remain_s.html`. Accessed: 2017-06-15.

[27] Y. Sun, A. Edmundson, L. Vanbever, O. Li, J. Rexford, M. Chiang, and P. Mittal. RAPTOR: Routing Attacks on Privacy in Tor. In *Proceedings of the 24th USENIX Security Symposium (USENIX Security 2015)*, pages 271–286. USENIX Association, 2015.

[28] The Tor Project. FAQ. `https://www.torproject.org/docs/faq.html.en`. Accessed: 2017-06-15.

[29] The Tor Project. Tor security advisory: Old Tor Browser Bundles vulnerable. `https://blog.torproject.org/blog/tor-security-advisory-old-tor-browser-bundles-vulnerable`. Accessed: 2017-06-15.

[30] The Tor Project. Understanding and Using Tor - An Introduction for the Lay(wo)man. `https://trac.torproject.org/projects/tor/wiki/doc/TorALaymansGuide`. Accessed: 2017-06-15.

[31] R. Wash. Folk Models of Home Computer Security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security (SOUPS 2010)*, pages 11:1–11:16. ACM, 2010.

[32] R. Wash and E. Rader. Too Much Knowledge? Security Beliefs and Protective Behaviors Among United States Internet Users. In *Proceedings of the Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 309–325. USENIX Association, 2015.

[33] P. Winter and S. Lindskog. How the Great Firewall of China is Blocking Tor. In *Proceedings of the 2nd USENIX Workshop on Free and Open Communications on the Internet*. USENIX Association, 2012.

[34] B. Zantout and R. Haraty. I2P data communication system. In *Proceedings of the 10th International Conference on Networks (ICN 2011)*, pages 401–409. ACM, 2011.

[35] M. E. Zurko and R. T. Simon. User-centered Security. In *Proceedings of the 1996 Workshop on New Security Paradigms (NSPW 1996)*, pages 27–33. ACM, 1996.

APPENDIX

A. SCREENING QUESTIONNAIRE

We invite you to participate in our study. Your participation will benefit science and help us understand user perceptions of software.

Our study involves a one-on-one interview. You may participate in-person or remotely via telephone or Voice-over-IP solutions, such as Skype. The interview will take a maximum of 45 minutes. Each participant will be compensated with a $20 Starbucks gift card.

To register, please answer the brief questionnaire below. We will contact you if we have an available position. Slots are limited, so if you wish to participate, please sign up as soon as possible.

If you have any questions, please contact us via email.

1) Age

- 18-24
- 25-34
- 35-44
- 45-54
- 55+
- Prefer not to say

2) Gender

- Male
- Female
- Other
- Prefer not to say

3) Email
Please enter your email address. This is the email address we will use to contact you.

4) Which of the devices below do you own and use? (Check all that apply.)

- Desktop Computer
- Laptop Computer
- Smartphone
- Tablet
- Other

5) Which of the technologies and services below have you ever used? (Check all that apply.)
[NOTE: Options were presented in random order.]

- Social Networking (Facebook, Twitter, LinkedIn, etc.)
- Online Audio and Video Conferencing (Skype, Facetime, etc.)
- Anonymization Software (Tor, etc.)
- Office Software (Word, Excel, Powerpoint, etc.)
- Online Music, TV, and Media
- Version Control Software (Git, Subversion, etc.)
- Online File Sharing (Dropbox, OneDrive, etc.)
- Mobile Messaging (Kik, Telegram, Snapchat, etc.)
- Online Banking
- Encryption Software
- Online Communities (Reddit, etc.)
- Computer Programming
- Online Shopping
- Blogging

B. INTERVIEW PROTOCOL

Thank you for taking the time to participate in this interview. The purpose of this interview is to discover your views, opinions, and understanding regarding how Tor works. Many people use Tor everyday for many reasons, from reading their email to accessing blocked Web sites. Please keep in mind that there is no single correct answer to these questions. Please answer the questions based on your own knowledge and experiences.

1. What do you do for a living? What does that entail?
2. What kind of computer(s) or mobile device(s) do you use? What are the differences (if any) in what these device(s) can do and how you use them?
3. On which of these device(s) do you use Tor?
4. When did you start using Tor? Why did you start using Tor?
5. How did you discover Tor?
6. Why do you currently use Tor?
7. This is a drawing exercise. Keeping background processes in mind, please draw what happens when you use the Tor Browser Bundle. Also note of who can access information about you. Please think aloud and explain your thought process while you are drawing.
8. How often do you use Tor?
9. What other browsers do you use?
10. Under which circumstances do you use the Tor Browser Bundle instead of another browser or vice versa?
11. Describe your feelings regarding the advantages and disadvantages of using Tor.
12. In what ways, if any, do you use Tor differently on your mobile device(s) than your computer(s)? (If applicable.)
13. Please fill out the given table of tasks and various entities involved in those tasks. For each of the tasks, mark the entities that you believe can access information about you when you perform the task using Tor. Please also mention what information you believe they can access. (See Table 1.)
14. Currently, a debate is going on about the role of privacy tools in matters pertaining to national security. Some people claim that strong privacy tools like Tor are good, while others claim they are bad. This is a part of a larger discussion about the trade-off between privacy and national security concerns. What is your opinion on this matter?
15. Is there anything else you would like to tell us? Is there anything that we should have asked?

Privacy Expectations and Preferences in an IoT World

Pardis Emami-Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer,
Lorrie Faith Cranor, Norman Sadeh
Carnegie Mellon University
Pittsburgh, PA, USA
{pardis, srutib, lbauer, lorrie}@cmu.edu
{htq, degeling, sadeh}@cs.cmu.edu

ABSTRACT

With the rapid deployment of Internet of Things (IoT) technologies and the variety of ways in which IoT-connected sensors collect and use personal data, there is a need for transparency, control, and new tools to ensure that individual privacy requirements are met. To develop these tools, it is important to better understand how people feel about the privacy implications of IoT and the situations in which they prefer to be notified about data collection. We report on a 1,007-participant vignette study focusing on privacy expectations and preferences as they pertain to a set of 380 IoT data collection and use scenarios. Participants were presented with 14 scenarios that varied across eight categorical factors, including the type of data collected (e.g. location, biometrics, temperature), how the data is used (e.g., whether it is shared, and for what purpose), and other attributes such as the data retention period. Our findings show that privacy preferences are diverse and context dependent; participants were more comfortable with data being collected in public settings rather than in private places, and are more likely to consent to data being collected for uses they find beneficial. They are less comfortable with the collection of biometrics (e.g. fingerprints) than environmental data (e.g. room temperature, physical presence). We also find that participants are more likely to want to be notified about data practices that they are uncomfortable with. Finally, our study suggests that after observing individual decisions in just three data-collection scenarios, it is possible to predict their preferences for the remaining scenarios, with our model achieving an average accuracy of up to 86%.

1. INTRODUCTION

The Internet of Things (IoT), composed of network-connected physical objects, is growing rapidly. The devices that make up the IoT vary greatly in their form and purpose, from sensors that people voluntarily carry on their wrists, to network-connected thermostats, to street lights that count the number of people who pass by. While these devices bring about new services, increase convenience, and improve efficiency, they also bring privacy and security risks.

To fully realize the potential of IoT, individuals need to be sufficiently knowledgeable and aware to make informed decisions. Thus, IoT devices need to inform their users about their data collection practices and offer privacy choices that respect individual privacy preferences. Gaining traction on this problem requires nuanced understanding of societal norms and context, as well as individual needs [31, 35]. For example, most people tacitly accept being recorded on cameras and CCTV outdoors in public spaces, but express disdain for installing video surveillance systems inside the walls of their homes. As more complex IoT scenarios become possible, many other factors may play a role in determining individuals' privacy preferences. While some may feel comfortable with their location being tracked for the purpose of traffic prediction, they may consent to tracking only their work commute. Others may consent only if they are assured that their location data is retained and used in an anonymized form.

We conducted a large-scale online vignette study to identify the contribution of different factors (such as the type of data, retention time, purpose of data collection, and location of data collection) in promoting or inhibiting individuals' self-professed comfort levels. We also studied the factors that trigger a desire for notifications about data collection. Our research identified which aspects of data collection or use by various IoT devices are most likely to cause discomfort, how realistic participants think these scenarios are, and which aspects they would like to be made aware of.

The results of our study informs the design of more transparent IoT-connected systems—we envision our results can be used to improve privacy notices for IoT devices, and develop more advanced personal privacy assistants [25].

This paper makes two main contributions. First, we show that individuals' comfort levels in a variety of IoT data collection scenarios are related to specific aspects of that data collection. Many of our findings are consistent with observations made in prior work, but our quantitative methodology and the scale of our experiment allows us to understand the effect of individual factors and their relative importance more precisely. Second, leveraging our qualitative and quantitative results, we advance explanations for many of the differences among these factors. We show that whether or not participants think the use of their data is beneficial to them has a profound influence on their comfort level. We also find that participants' desire for notification is closely related to whether or not they feel comfortable with data collection in a particular scenario.

The paper is organized as follows. First, we discuss related work. Then we describe the design of our vignette study, and discuss our quantitative and qualitative analysis of our survey data. Next, we present the results of our prediction model, and draw conclusions from the analysis. Finally, we discuss study limitations and possible approaches to mitigate some of the concerns highlighted by our study.

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.

2. RELATED WORK

Our research builds on prior work aimed at understanding individuals' IoT-related privacy concerns, and potential solutions for mitigating them [6, 8]. Additionally, prior research has studied various factors that can impact privacy preferences, the results of which were used to inform the design of our study. Recent work has also developed models to predict individuals' privacy preferences, so that data collection can be personalized to suit people's privacy preferences. Our work aims to address privacy concerns in a variety of IoT scenarios where sensing is pervasive. Our work underscores the relative importance of different privacy concerns to individuals. These findings inform the understanding of privacy preferences as they relate to IoT data collection.

2.1 IoT Privacy Challenges

New methods of data collection in the IoT have led to new privacy challenges. Some of these challenges include obtaining consent for data collection, allowing users to control, customize, and choose the data they share, and ensuring the use of collected data is limited to the stated purpose [33]. These challenges are made more difficult by the increased potential for misuse of personal information in the IoT domain. This stems from the pervasive tracking of habits, behaviors, and locations over a long period of time. There are new risks to personal safety introduced by IoT systems [6, 9]. Addo et al. demonstrated that trustworthiness of an IoT application is impacted by the implemented privacy and security practices [2]. To be accepted by consumers, IoT-connected device developers must consider the privacy and security implications of their products.

2.2 Privacy Interfaces for IoT Systems

There have been several proposals to help address privacy concerns related to data collection in the IoT domain. Mehrotra et al. presented two systems that could help highlight privacy challenges associated with IoT sensing and allow for testing of various privacy-enhancing solutions [30]. Lederer et al. identified five "pitfalls" in designing systems, particularly in ubiquitous computing environments, which lead to negative implications for individual privacy [18]. To address some of these pitfalls, Egelman et al. used crowdsourcing techniques to study different designs of privacy icons for a camera, with the aim of helping individuals make an informed decision about their privacy. Though many of their icons were successful in conveying what data was being collected, many participants demonstrated low comprehension. These findings underscored the difficulty of successfully informing individuals about what is going on around them in an IoT setting [12]. Recognizing the privacy risk caused by involuntary disclosure of information in IoT environments, Ukil et al. proposed a privacy management scheme that estimates a domain-specific measure of risk due to privacy disclosure in smart energy applications [38].

According to Bhaskar et al., a major limitation of prior work studying privacy in IoT environments is that studies typically focus on a single environment in which IoT sensing is occurring [6]. Thus, many of the proposed solutions do not generalize to other IoT contexts. Our work attempts to address this shortcoming by identifying privacy concerns in multiple heterogeneous scenarios which employ different types of data collection. This way, our methodology can determine which factors have the greatest impact on measures of individuals' comfort with data collection. The results can inform the design of privacy-enabling solutions appropriate to the variety of contexts we have studied. Furthermore, our study aims to expand beyond prior work in this area by identifying privacy concerns individuals have in data collection scenarios which are not obviously aligned with specific privacy risks.

2.3 Factors Impacting Privacy Preferences

Prior studies outside of the IoT context have examined different factors that can impact individuals' willingness to share information, based on measures of comfort with data collection. Bilogrevic et al. found that the comfort levels associated with sharing data are highly dependent on the specific type of data and the sharing context (e.g. search engines, social networks, or online shopping sites) [7]. Leon et al. tested whether data retention, access to collected information, and the scope of use affected willingness to share data for online behavioral advertising purposes. Individuals were more willing to share certain types of data if it had a retention period of one day, but for periods longer than one week, individuals were less likely to be willing to share [22].

Other work has focused on privacy preferences related to mobile devices and applications. Lin et al. evaluated individuals' perceptions of requests to access privacy-sensitive resources (e.g. sensors) on mobile devices. They found that both individual expectations of what an app does and the purpose for which an app requests access to sensitive resources impacts their privacy decisions [23]. In order to better understand people's attitudes toward sharing their location in mobile applications, Sadeh et al. built a system that enabled mobile device users to select and limit with whom they want to share their location. They concluded that increasing people's awareness has a critical role in helping them define more precise policies for protecting their privacy [36]. Tsai et al. studied the impact of giving feedback to mobile device users. Their study informed participants about who their data is being shared with, and when the data was shared. The goal was also to help people manage their privacy on a location sharing application. They reported that when people get adequate feedback, they are more willing to share data. They were also more comfortable with sharing their location [37].

Other studies more closely aligned with our work have evaluated several factors that may impact privacy concerns related to IoT data collection. Lederer et al. studied the relative importance of two factors; the entity collecting data, and the situation in which it is being collected, for determining users' privacy preferences in ubiquitous computing settings. Their results indicate that individuals base their privacy decisions on who is collecting their data, rather than the context in which it is being collected [19]. Lee and Kobsa tested five factors related to the context of data collection in two separate studies and found that individuals generally thought that monitoring in personal spaces was unacceptable, along with monitoring by an unknown entity or the government. Their results also indicate that photo and video monitoring may cause some privacy concern regardless of context [20, 21]. Other small, qualitative studies have focused on individuals' privacy preferences related to wearable sensors. These studies revealed that people demand ownership of the data they produce, and that privacy concerns vary depending on factors including retention time and the perceived value of the data collected [4, 17].

Our work leverages prior work to identify several factors that may impact individuals' privacy concerns and preferences in IoT settings. While data retention was found to be a significant factor in an online context [22], we aim to determine whether this remains true for IoT data collection. Additionally, the impact of the location of the data collection, type of data being collected, and purpose for collection have already been studied in prior work considering IoT contexts [20, 21]. We aim to expand on these findings by evaluating these factors in a larger scale study, and in combination with additional factors capturing more contextual nuances that are specific to IoT environments.

2.4 Predicting Privacy Preferences

Prior work has shown that privacy preferences can be inferred by segmenting collections of individuals based on profiles. These profiles represent clusters of different individuals and their privacy decisions. In the mobile app privacy domain, Lin et al. and Liu et al. demonstrated that a small number of profiles may be capable of predicting individuals' decisions to allow, deny, or be prompted for app permissions with a high level of accuracy [24, 26]. In IoT data collection scenarios, Lee and Kobsa were able to identify four clusters of participants with distinctive privacy preferences. These clusters were used to predict their study participants' decision to allow or deny monitoring in a particular IoT context with 77% accuracy [21]. In our work, we incorporate additional factors into a larger scale study, using similar techniques to make predictions with the goal of achieving improved prediction accuracy relative to prior work.

3. METHODOLOGY

We conducted a within-subjects survey with 1,014 Amazon Mechanical Turk [1] workers in order to understand individuals' privacy preferences. We exposed each participant to 14 different vignettes presenting an IoT data collection scenario. Vignettes are "short stories about hypothetical characters in specified circumstances, to whose situation the interviewee is invited to respond," [13] and have been used in prior work studying varying privacy contexts [28, 29].

Between vignettes, we varied eight factors that we hypothesized could influence individuals' privacy preferences:

- the type of data collected (data_type),
- the location where the data is collected (location),
- who benefits from the data collection (user_benefit),
- the device that collects the data (device_type),
- the purpose of data collection (purpose),
- the retention time (retention),
- whether the data is shared (shared), and
- whether additional information could be inferred from the collected data (inferred).

Several of these factors have already been shown in prior work to be important to individuals, when presented individually or in combination [4, 17, 19, 20, 21, 22]. Our design allowed these factors to be studied simultaneously, capturing more contextual nuances. In our vignettes, some factors could take on one of many possible levels. For reference, table 1 describes the factors and their corresponding levels.

After accepting the MTurk HIT, each study participant was directed to a survey where they were shown 14 different vignettes.

Each vignette introduced the factors being tested in the same order. In each scenario, vignettes began with the location of the data collection and ended with the retention period. The following is an example of a scenario presented to participants:

> You are at **work** and your **smart watch** is keeping track of your **specific position in the building**. Your position is shared with the **device manufacturer** to **determine possible escape routes in the case of an emergency or a hazard**. This data will be kept by the manufacturer **until you leave for the day**.

All factorial combinations of the different levels of each factor produced 126,720 possible scenarios, many of which contained combinations of factors which did not make sense (e.g. a presence sensor taking iris scans for emergency purposes). These scenarios were removed from the set of scenarios shown to participants. From the remaining set, we selected 380 scenarios that could feasibly occur, and ensured that this subset contained scenarios in which each level of each factor was represented. 14 vignettes drawn from these 380 scenarios so as to not overburden them. Randomly selecting subsets of 14 scenarios could have caused interaction effects due to a lack of diversity in each factor (e.g., presenting only one retention time on otherwise diverse scenarios) [3]. To minimize such interaction effects, we carefully selected subsets of vignettes so that every level of every factor was present at least once per subset, with the exception of the factors device_type, purpose, and inferred, which were dependent on other factors such as location, device_type, and user_benefit. In doing so, we divided the list of scenarios into 39 subsets with 14 scenarios each, and presented each participant with vignettes corresponding to one of these 39 subsets. The subsets were not mutually exclusive.

For each scenario, participants were asked how comfortable they were with data collection in that scenario and whether they found the use of data in the scenario to be beneficial (user_perceived_benefit). This factor is different from user_benefit, which refers to whether the data collection benefits the participant or the collector and is part of the scenario design; user_perceived_benefit refers to the participant's perception of whether the scenario would be beneficial to them. This question was only asked about scenarios in which a purpose was given; we coded this factor as 'N/A' for scenarios without a purpose. We also asked participants whether they would allow the data collection described in the scenario, and how often they would like to be informed about the data collection. Further questions asked how realistic a scenario was ("I think scenarios like this happen today," "... will happen within 2 years," and "... will happen within 10 years") and coded the answers to these three questions as happening_today, within_two_years, and within_ten_years, respectively. These three questions were answered on a five-point Likert scale from "Strongly Disagree" to "Strongly Agree" and were binned into binary categories based on agreement—0 (strongly disagree, disagree) and 1 (strongly agree, agree, neither agree nor disagree). Finally, we asked participants general demographic questions, followed by ten questions from the Internet Users' Information Privacy Concerns (IUIPC) scale to gauge their level of privacy concern. The IUIPC scale questions focus on concerns about control, awareness, and collection [27]. The complete set of questions asked in our survey is included in the Appendix.

3.1 Factors Impacting Preferences

We were interested in learning what factors of data collection contributed most significantly to individuals' comfort and preferences. Thus, we asked questions about how comfortable they were with the given scenario. We also asked if they would allow a specific data collection or not, and how often they would want to be notified about it. Participants' responses to these questions enabled us to build models that predict the concerns and preferences of the general population, based on our sample. We constructed five statistical models, capturing five dependent variables: comfort level, allow or deny decisions for the data collection, desire to be notified of data collection every time, desire to be notified once in a while, and desire to be notified only the first time. In addition to the eight factors in Table 1, we included the factors user_perceived_benefit, happening_today, within_two_years, within_ten_years, gender, age, income, and education, as well as the three IUIPC scale factors IUIPC-control, IUIPC-awareness, and IUIPC-collection.

[1] Amazon's Mechanical Turk `https://www.mturk.com`

Factor	Levels	Description
location	department store; library; workplace; friend's house; home; public restroom	location where the data is collected
data_type	presence; video; specific position; biometric data (e.g., fingerprint, iris, face recognition)	type of data collected
device_type	smart watch; smart phone; camera; presence sensor; temperature sensor; fingerprint scanner; facial recognition system; iris scanner	device that is collecting the data; some devices like smart phones can collect multiple data types
user_benefit	user (e.g., get help in emergency situations); data collector (e.g., downsize staff)	who benefits from the data collection and use
purpose	a specific purpose is mentioned; it is mentioned that participants are not told what the purpose is	purpose of data collection depends on the location, the data and who is benefiting
retention	forever; until the purpose is satisfied; unspecified; week; year	the duration for which data will be kept
shared	shared (e.g., with law enforcement); no sharing is mentioned	whether the data is shared or not
inferred	inferred (e.g., movement patterns); inferred data is not mentioned	Additional information can be inferred and users can be deanonymized

Table 1: Factors varied between vignette scenarios, levels of the factors presented in scenarios, and description of each factor.

We represented income as a quantitative variable based on categories of income ranges, excluding two outliers—participants who reported earning more than $200,000. We mapped all Likert scale responses to binary categories of 0 and 1, where 1 implies a positive preference, and 0 implies a negative preference. All of the quantitative variables (income, age, IUIPC-control, IUIPC-awareness, IUIPC-collection) were normalized before analysis to be on the same scale with a mean of 0 and standard deviation of 1.

We did not include two of the eight privacy factors, device_type and purpose. The device that is collecting the data was mentioned in the vignettes to make them more realistic, but was not considered in the statistical analysis because the device was uniquely determined by the type of data that was collected. The type of data that was collected was considered in the statistical analysis, resulting in a dependency between the two factors. Dependencies of this type between factor levels can lead to inaccurate statistical inferences. To improve the accuracy of our results, we excluded them from our statistical analysis. For the same reason, we removed purpose as it was not linearly independent from multiple other factors, such as location and user_benefit. Treating it as an independent factor would have resulted in scenarios that did not make sense contextually. For instance, using purpose as an independent factor would have included scenarios which involved collecting fingerprints to downsize staff. To eliminate these nonsensical scenarios from our study, we chose to remove purpose from the analysis, instead of the other factors on which it depended.

After removing these two factors, we found one of the subsets of scenarios contained two scenarios that differed only in these two factors. Therefore, for participants who received this subset, we removed the first of the two scenarios' answers and analyzed the remaining 13 scenarios.

Our models were constructed using generalized linear mixed model (GLMM) regression with a random intercept per participant. GLMM is particularly useful for modeling repeated measures experiments, such as ours, in which participants are presented with multiple parallel scenarios [5].

We performed model selection to find the best combination of factors by using a search algorithm with a backwards elimination approach. For each of our dependent variables, we found the model that best fit the data according to the Bayesian Information Criterion (BIC). We eliminated the variables with the largest p-value in each step of the model selection and continued the elimination until the BIC reached the global minimum [15]. The model with the lowest BIC best explains the dependent variable.

We present the regression tables for our best models in the Results section. We used a significance threshold of 0.05 to determine whether or not a factor was significant. Effects and the effect size of a factor level can be interpreted as proportional to the magnitude of the estimate co-efficient. We also defined a baseline for each factor. The regression tables and co-efficients of levels in the model were computed against the corresponding factors' baseline. Some of the baselines were selected based on specific concerns highlighted by our qualitative data, such as data_type (baseline = specific position) and location (baseline = friend's house). The baselines for other factors were selected based on their alphabetical ordering.

3.2 Predicting Preferences

Using the results from the model selection for each dependent variable, we further examined their predictive ability for individuals' preferences. Specifically, in our analysis we focus on predicting:

- an individual's comfort with a specific data collection scenario; and
- an individual's decision to allow or deny a specific data collection instance.

We believe that the ability to predict individuals' preferences or decisions is useful, since we can imagine deployment scenarios where a system needs to predict an individual's comfort or decision to allow or deny data collection. In these cases, the system would have more data accumulated over time specific to an individual using the system, and so would likely perform better than the classifiers in our experiments.

3.2.1 Features

For each of the two prediction tasks mentioned above, we used the main factors and interactions from the results of our model selection to predict the two outcomes; comfort level, and the decision to allow or deny.

Continuous features were encoded as-is in the feature vector, while categorical features were encoded as one-hot vectors for each category in the domain of that feature. This means, that each categorical variable was encoded as a vector of binary features where each feature corresponded to the binary value of one of the categories in the original categorical variable. In a one-hot vector, only one value in the whole vector will be 1 at any given time. This is a common way of encoding multi-class categorical features for machine learning tasks. For each categorical variable, the overall feature vector was

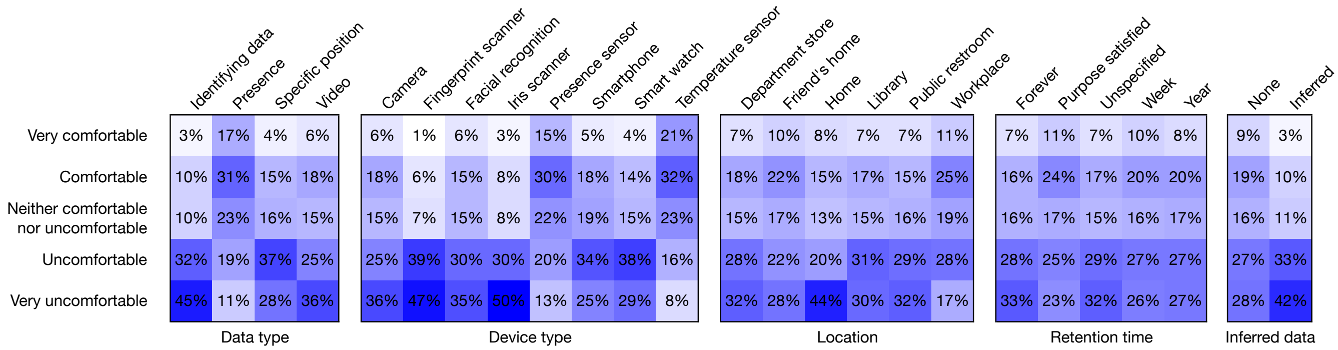


Figure 1: Summary statistics showing the relation between various factors and participants' comfort level. For example 45% of participants were very uncomfortable when the type of data being collected was biometric. Cells with larger numbers are darker in background color.

Gender		Age		Education		Income		IUIPC Score	
Male	49.2% (49.2%)	Range	18-78	No high school	0.8% (10.9%)	< $15k	16.4% (11.6%)	*Control Factor*	
Female	50.1% (50.8%)	Mean (SD)	36.1 (10.9)	High school	30.8% (28.8%)	$15k-$34k	33.8% (20.5%)	Range	1.33-7
No answer	0.7% (0.0%)	US average	37.9	Associates	9.7% (10%)	$35k-$74k	36.1% (29.4%)	Mean [SD]	5.95 [0.90]
				Bachelors	49.0% (48.7%)	$75k-$149k	9.3% (26.2%)	*Awareness Factor*	
				Professional	8.5% (1.5%)	$150k-$199k	0.9% (6.2%)	Range	1-7
				No answer	1.0% (0.0%)	> $200k	0.2% (6.1%)	Mean [SD]	6.44 [0.82]
						No answer	3.2% (0.0%)	*Collection Factor*	
								Range	1-7
								Mean [SD]	5.79 [1.11]

Table 2: Demographic breakdown of our participants. In the Gender, Education, and Income columns, the numbers in parentheses show the US average, according to census data from 2015.

increased in size by the size of each one-hot vector. For interactions between whole factors, we computed the product of each combination of the values in the one-hot vector and appended this vector of interaction products to the feature vector.

3.2.2 *Classifiers*

We experimented with various binary classifiers for the allow/deny prediction, and both binary and continuous classifiers for the comfort prediction. For binary classifiers where the outcome is binary, we used logistic regression, support vector machines (SVM), k-Nearest Neighbor, AdaBoost (with various weak base classifiers), and simple neural networks in the form of three-layer multi-layer perceptrons (MLP) [32]. For predicting comfort, we also experimented with a continuous version of the comfort level on a scale from 1 to 5, normalized to be between 0 and 1, for which we used linear regression for prediction.

We found the AdaBoost classifier with a logistic regression base classifier (with $l2$-regularization) to be the best performing, and these are the results we report on. We implemented our classifier and ran experiments using the Scikit-learn Python library [32].

3.2.3 *Evaluation Methodology*

We tested using two different sizes of the training data for predicting a specific participant's preferences: 75% of 100% of the answers provided by the remaining participants. In all cases, training data also included the participant's own answers to three of the scenarios they were asked about; we tested on the remaining 11 scenarios (10 scenarios in the case of the participants mentioned in Section 3.1).

When predicting comfort level, we report accuracy in two ways, which differ in how they treat predictions when the participant did not have a preference. In the first approach, we counted any prediction as correct if the participant's actual survey response fell in the middle of the Likert scale, i.e., their answer was "Neither Agree nor Disagree." We did this based on the reasoning that if an individual doesn't have an explicit preference, then any prediction would be consistent with that preference. In the second approach, we report accuracy by testing only on scenarios for which a participant did not answer neutrally. This measures how many of a participant's non-neutral preferences can be predicted.

Additionally, for both prediction tasks, we report the results of using a simple majority classifier that classifies each element in the test set as the majority class within the training set.

In each experiment, we randomly selected 50 participants whose answers to predict. We report the accuracy, precision, and recall of the classifier averaged over the 50 participants.

Accuracy is the fraction of predictions that were accurate. Both precision and recall are indicators for measuring the effectiveness of a classifier in predicting positive examples. For predicting comfort, a positive example is a scenario for which the user's answer falls into the "comfortable" category. For predicting allow/deny decisions, a scenario for which a user answers "Allow" is a positive example. *Precision* is the fraction of positive predictions during testing that are actually correct according to the training data. *Recall* is the fraction of all positive examples in the training data that the classifier predicts as positive during testing.

For each participant, we used a form of cross-validation defined as follows:

For $X = 75\%$ or $X = 100\%$ of training data:

- Randomly select 50 participants as targets for prediction.
- For each participant, run 6 different iterations of prediction.
- In each of the 6 iterations, randomly select $X\%$ of training data from the remaining participants and randomly select 3 responses from the total set of scenarios the target was asked

Categories	Tags (Usage)	Examples
Factors (n = 842)	purpose (63%), data (26%), retention (25%), sharing (18%), benefit (17%), location (7%), device (2%),	P880:"It would make me more comfortable knowing where this data was going and how it was going to be used, as well as it being consented."
Whitelist (n = 350)	safety (42%), anonymous_data (40%), personal_benefit (7%), public (7%), common_good (6%), improve_services (6%)	P908: "If they helped to make me safer in some way.", P779:"I'd be fine with data that doesn't identify me.", P121:"That my safety was the reason for it, or saving me money"
Blacklist (n = 474)	biometrics (26%), personal_information (20%), everything (16%), location (13%), private_location (12%), bathroom (9%), video (9%), commercial (8%), government (6%), law_enforcement (5%)	P136:"[..] that they might share the data with other parties [..]. Also, knowing that a retinal or fingerprint scan might be stolen and used to gain access to something else." P415:"The government spying on me in my home, or private corporations using that data to identify me [..], no way."
Information (n = 417)	purpose (66%), retention (35%), sharing (21%), collector (15%), access (13%), data_handling (13%), data_security (5%)	P271: "Knowing exactly what the data is used for, where it is stored, who it is shared with, and when it is collected."
Control (n = 113)	deletion (33%), consent (30%), opt-out (27%), ownership (14%), access (13%), copying (10%)	P913: "Nine times out of ten I won't care and would be happy to allow it, I just want to be informed and have the ability to deny consent should I choose."
Risks (n = 298)	misuse (29%), surveillance (18%), data_security (18%), privacy (16%), tracking (12%), intransparency (8%),	P286:"I don't want my personal information getting into the wrong hands." P47:"I don't like the idea of government organizations being alerted of my location at all times."

Table 3: Categories and codes used to code free text answers. Percentages in brackets are the number of times a code was used when the category was coded, multiple codes could be applied per category. Rows on Factor/Whitelist/Information/Control refer to answer to the question "..what would make you uncomfortable with sharing data in such situations?" Blacklist/Risks stem from the answers to the question about discomfort.

about. This data is used for training; testing is done on the remaining scenarios of the target.

- Calculate the average accuracy, precision, and recall scores averaged over 6 iterations each and over the 50 random participants.

We report on the results of our experiments in Sections 4.2.2 and 4.3.2.

3.3 Qualitative Analysis of Preferences

We also qualitatively analyzed participants' responses to the free-response questions they were asked at the end of the survey. The answers were coded with regards to five topics: the factors that were mentioned; whether specific scenarios were described as comfortable or uncomfortable; what the participant wants to be informed about; and what means of control (e.g. access, edit, ability to delete) they request. A codebook was developed from 100 answers and applied to another set of 100 answers by two annotators independently. They reached an inter-annotator agreement of 0.89 (Cohen's Kappa) for whether a topic was addressed and between 0.67 and 0.72 on the actual tags (e.g., which factor was mentioned). After achieving this accuracy, the remaining answers were divided among the two annotators and coded by one annotator each. A summary of categories and codes and their occurrence is shown in Table 3.

4. RESULTS

In this section, we describe our participants and present results regarding participants' comfort level with different data collection scenarios, their decisions to allow or deny data collection, and desire to be notified.

4.1 Participants

Our survey was completed by 1,014 MTurk workers. We removed the answers of seven participants because they took less than five minutes to complete the survey, while the average completion time was 16 minutes. This resulted in 1,007 participants whose responses we included in our analyses. Participants were required to be from the United States and have a HIT approval rate of above 95%. Table 2 describes participants according to their demographics and privacy concern level. Our participants were slightly better educated and had a higher income than the U.S. average.

4.2 Comfort with Data Collection

In our survey, after presenting each scenario we asked: "How would you feel about the data collection in the situation described above if you were given no additional information about the scenario?" We measured participants' comfort on a five point Likert scale from "Very Comfortable" to "Very Unfomfortable" with the middle point of "Neither Comfortable Nor Uncomfortable."

Figure 1 shows the general distribution of participants' comfort across different levels of each factor. Participants were strongly uncomfortable if the scenarios they were asked about had biometric as data_type (45% strongly uncomfortable), device_type as iris scanner (50% strongly uncomfortable), location as their home (44% strongly uncomfortable), retention as forever (33% strongly uncomfortable), or if other data was inferred from the data collection (42% strongly uncomfortable).

4.2.1 Factors Impacting Comfort Level

Using the best model, we ordered the factors based on their contribution to comfort level by looking at the change in BIC when each factor was added to the null model (the model that has no factor other than random intercept for participants). Table 4 shows the factors ordered by their effect sizes from the most effective factor (the interaction between the data_type and happening_today) to the factor with the lowest effect size (retention). As shown in the table, not all levels of the factors are statistically significant ($p < 0.05$). A positive estimate (effect size) indicates inclination toward comfort and a negative estimate shows inclination toward discomfort.

Scenarios in which video was being collected and participants thought such data collections are happening_today had the greatest positive impact on participant comfort with data collection ($p < 0.05$, coefficient = 1.38). This is in line with our qualitative results, where we found that 38% of all participants mentioned a specific scenario with which they were comfortable (category "whitelist," Table 3), and from the whitelisted scenarios, 42% mentioned safety, security, or emergency situations as specific purposes for data collection that they would generally approve of. Another 40% of those who whitelisted a scenario were less concerned when anonymous or anonymized data was involved. When an example was given, participants mentioned scenarios involving presence or temperature sensors as ones they would be comfortable with.

Scenarios in which biometric information (e.g., fingerprint, iris image) was being collected and participants thought such data collection is happening_today, had the greatest negative impact on participant comfort ($p < 0.05$, coefficient = 0.89). This is also in line with our qualitative analysis of answers to the question "Keeping in mind the 14 scenarios, what would make you uncomfortable with sharing data in such situations?" In 46% of the answers, participants conveyed one or more specific things that they did not want to happen (coded in category "blacklist," Table 3). Within these answers, the collection of biometric data_type was mentioned by 26%.

Based on previous findings [7], we hypothesized that participants would be less comfortable if a scenario included the explicit notice that collected data would be shared with others (shared). Consistent with that hypothesis, we found that informing participants that data would be shared with third parties (e.g., with the device manufacturer or law enforcement) caused participants to be less comfortable ($p < 0.05$, coefficient = -0.68). The qualitative results show that a minority of participants expressed mistrust of or discomfort with sharing with government (6%) and law enforcement (5%) agencies.

Within the qualitative responses related to discomfort, we also found explanations of why participants did not want to share their data. About 29% of all participants mentioned some perceived risk, ranging from the fear of identity theft or the use of data for other than the stated purpose (misuse) to a general concern about privacy and surveillance in general. Among those that mentioned a perceived risk, 29% feared that their data could be used in a way that would harm them or put them at a disadvantage. About 18% of these answers explicitly mentioned data security issues and leaks as a cause of concern.

> P11: [I'm concerned about] any unique identifiers that could be hacked and then used for identity theft, blackmail, humiliation, etc.

With respect to the location of data collection, most levels had small, positive effect on comfort level. As described above, only scenarios taking place at home had a negative impact on the perceived comfort. Our qualitative results further substantiate this, as participants who mention location as a factor that made them comfortable often cited the dichotomy between public and private places. Data collection in private places is described as highly intrusive while data collection in publicly accessible spaces like libraries or stores was described as "ok." Out of the 474 participants that expressed discomfort with specific scenarios, those that took place in one's home (12%) and in bathrooms (8%) were most frequently mentioned.

The factor retention had the smallest effect size on the results and only short retention times (immediate deletion or storing for a week) had a significant, positive effect on the comfort level. This is in line with the qualitative results were, about 25% of those that mentioned a specific factor in their answers referred to how long their data was stored. Those that explicitly mentioned a time span favored a retention time of less than a week.

4.2.2 Predicting Comfort Level

As explained in Section 3, we trained a machine learning model to predict a participant's comfort based on the significant factors and interactions determined through model selection. The results are shown in Table 5.

The classifier achieved an average accuracy of around 81% over 50 different participants when either 100% or 75% of the other participants' answers are used as training data.

There is a sizable difference in precision and recall depending on whether (1) predictions are counted as correct whenever participants expressed neither a positive nor a negative opinion or (2) scenarios in which participants did not express an opinion are removed from the test data. As per the discussion in Section 3.2.3, both ways of measuring performance are indicative of the utility of using a similar classifier in practice.

Factor	Estimate	Std Err	Z-value	p-value	BIC
data type:happening today					14633
baseline=friend's house:not happening today					
video:happening today	1.39	0.20	6.83	**0.00**	
biometric:happening today	0.89	0.15	5.80	**0.00**	
presence:happening today	0.91	0.18	12.57	**0.01**	
temperature:happening today	0.95	0.22	4.26	**0.00**	
data (baseline=specific position)					15843
biometric	-1.45	0.13	-11.12	**0.03**	
presence	1.42	0.16	8.99	**0.00**	
temperature	2.50	0.20	12.57	**0.00**	
video	-0.30	0.19	-1.62	0.11	
user perceive benefit:location					15866
baseline=beneficial:friend's house					
not beneficial:department store	0.00	0.32	0.00	0.99	
purpose unspecified:department store	-0.07	0.24	-0.30	0.76	
not beneficial:house	-0.15	0.48	-0.30	0.76	
purpose unspecified:house	0.05	0.28	0.19	0.85	
not beneficial:library	-0.45	0.33	-1.38	**0.00**	
purpose unspecified:library	-0.17	0.24	-0.70	0.48	
not beneficial:public restroom	-0.40	0.36	-1.10	0.27	
purpose unspecified:public restroom	-0.48	0.26	-1.85	**0.01**	
not beneficial:work	-0.49	0.36	-1.38	0.17	
purpose unspecified:work	-0.11	0.24	-0.47	0.63	
being shared:user perceived benefit					15969
baseline=not being shared:beneficial					
being shared:not beneficial	-0.71	0.19	-3.70	**0.00**	
shared:purpose unspecified	0.37	0.13	2.94	**0.02**	
user perceived benefit (baseline=beneficial)					16055
not beneficial	-1.88	0.34	-5.60	**0.00**	
purpose unspecified	-1.30	0.25	-5.26	**0.04**	
retention:user perceived benefit					16058
baseline =unspecific:not beneficial)					
not deleted:not beneficial	-0.12	0.22	-0.06	0.96	
purpose specific:not beneficial	-0.30	0.28	-1.08	0.28	
week:not beneficial	0.49	0.23	2.11	**0.00**	
year:not beneficial	0.10	0.24	0.39	0.69	
not deleted:purpose unspecified	-0.43	0.16	-2.69	**0.00**	
week:purpose unspecified	-0.29	0.16	-1.76	0.07	
year:purpose unspecified	-0.22	0.17	-1.31	0.19	
happening within 2 years (baseline=disagree)					16199
agree	0.96	0.11	9.01	**0.00**	
happen today (baseline=disagree)					16491
agree	10.98	333.4	0.03	0.97	
location (baseline=friend's house)					17987
library	1.00	0.18	5.54	**0.00**	
work	0.87	0.18	4.82	**0.01**	
house	-0.88	0.20	-4.34	**0.00**	
department store	0.76	0.18	4.24	**0.00**	
public restroom	0.29	0.19	1.48	0.14	
being shared (baseline=not being shared)					18079
being shared	-0.68	0.09	-7.86	**0.00**	
IUIPC					
collection	-0.59	0.05	-11.47	**0.04**	18081
retention (baseline=not specified)					18103
week	0.25	0.11	2.25	**0.00**	
year	0.16	0.11	1.45	0.14	
purpose specific	0.0.56	0.15	4.85	**0.02**	
not deleted	0.10	0.10	0.99	0.32	

Table 4: Generalized linear mixed model regression output for the comfort level model. A positive estimate (effect size) indicates inclination toward comfort and a negative estimate shows inclination toward discomfort. Factors are ordered by their contribution: the factor with the lowest BIC contributes most to explaining participants' comfort level.

Class.	Training	Neutral	Acc.	Prec.	Recall
ABC	100% (1,006)	correct	81.06%	73.86%	83.06%
ABC	100% (1,006)	excluded	77.53%	54.50%	63.49%
ABC	75% (755)	correct	81.79%	71.30%	78.34%
ABC	75% (755)	excluded	77.67%	54.48%	60.77%
SMC	100% (1,006)	correct	72.03%	71.33%	40.92%
SMC	100% (1,006)	excluded	67.96%	0%	0%

Table 5: Accuracy, precision, and recall of (1) **ABC**: the AdaBoost classifier (with logistic regression as the base learner) and (2) the **SMC**: simple majority classifier, for predicting a user's comfort level with an instance of data collection. "Training" indicates the fraction (and number) of non-test participants used to train the classifier. "Neutral" indicates whether predictions are always counted as correct if a participant didn't indicate a preference for that scenario ("correct") or whether such scenarios are removed from the test set ("excluded").

Class.	Training	Acc.	Prec.	Recall
ABC	100% (1,006 users)	79.09%	76.79%	82.32%
ABC	75% (755 users)	79.09%	76.79%	82.32%
SMC	100% (1,006 users)	52.58%	0%	0%

Table 6: Accuracy, precision, and recall of (1) **ABC**: the AdaBoost classifier (with logistic regression as the base learner) and (2) **SMC**: the simple majority classifier, for predicting a user's decision to allow or deny data collection. "Training" indicates the fraction (and number) of non-test participants used to train the classifier.

Table 5 also describes the performance of our simple majority classifier that uses all non-test participants' answers as training data. These results form a baseline for understanding the performance of the AdaBoost classifier. Although a majority classifier is correct about 70% of the time, AdaBoost additionally correctly predicts more than a third of the predictions that the majority classifier gets wrong.

4.3 Allowing or Denying Data Collection

4.3.1 Factors Impacting Allow/Deny Decisions

We found a set of factors that can explain participants' response to the question: "If you had the choice, would you allow or deny this data collection?" We again ordered factors with respect to their effect size. The interaction of data_type and location has the most impact while shared has the smallest effect. By looking at the coefficient of the levels within each factor we can claim that participants were most likely to deny data collection in scenarios in which their presence was being collected at their workplace. Also, knowing that the data was being shared had the least effect on their preference to deny a data collection. In this model a positive estimate shows likeliness to deny and a negative estimate shows the likeliness to allow a data collection scenario. The regression results are shown in Table 7.

Among the common statistically significant factor levels, the ones that made participants more likely to be comfortable with a data collection also made them more likely to allow the data collection. Many factors were in line between the two models of comfort level and allow/deny such as data_type, location, user_perceived_benefit, shared, retention, happening_today, and within_two_years. However, the best model that described participants' comfort level (Section 4.2) was not the same as the best model that described the desire of participants to allow or deny a data collection. For example, we found that the interaction between data_type and location was the most helpful factor in the allow/deny model,

Factor	Estimate	Std Err	Z-value	p-value	BIC
data:location					15232
baseline=specific position:friend's house					
biometrics:department store	1.58	0.24	6.38	**0.01**	
presence:department store	1.22	0.37	3.3	**0.00**	
temperature:department store	1.61	0.55	2.94	**0.00**	
video: department store	-0.99	0.21	-4.83	**0.00**	
presence: house	0.42	0.41	1.02	0.31	
temperature: house	0.23	0.42	0.54	0.58	
biometrics:library	1.16	0.23	5.01	**0.01**	
presence:library	1.55	0.37	4.1	**0.01**	
temperature:library	1.52	0.43	3.52	**0.00**	
video:library	-0.5	0.2	-2.46	**0.00**	
presence:public restroom	1.87	0.36	5.11	**0.00**	
temperature:public restroom	1.54	0.38	3.99	**0.00**	
video:public restroom	1.36	0.36	3.77	**0.00**	
presence:work	2.11	0.34	6.1	**0.03**	
temperature:work	1.66	0.39	4.29	**0.00**	
being shared:user perceived benefit					15297
baseline=not being shared:beneficial					
being shared:not beneficial	0.62	0.19	3.26	**0.00**	
shared:purpose unspecific	-0.27	0.12	-2.1	**0.04**	
retention:user perceived benefit					15352
not deleted:not beneficial	-0.147	0.226	-0.65	0.515	
purpose-specific:not beneficial	0.39	0.248	1.37	0.17	
week:not beneficial	-0.126	0.24	-0.52	0.6	
year:not beneficial	-0.17	0.24	-0.68	0.49	
not deleted:purpose unspecified	0.45	0.16	2.81	**0.02**	
week:purpose unspecified	0.76	0.16	4.52	**0.00**	
year:purpose unspecified	0.48	0.17	2.85	**0.01**	
user perceived benefit (baseline=beneficial)					15374
not beneficial	2.85	0.17	16.38	**0.00**	
purpose unspecified	1.67	0.17	9.92	**0.01**	
data:happening today					15525
baseline=friend's house:not happening today					
video:happening today	-1.39	0.22	-6.26	**0.00**	
biometric:happening today	-0.78	0.16	-4.89	**0.00**	
presence:happening today	-0.95	0.19	-5.02	**0.02**	
temperature:happening today	-0.9	0.23	-3.87	**0.00**	
happening within 2 years:benefit of scenario					15986
baseline=disagree:benefit to company					
agree: purpose unspecified	0.12	0.36	0.34	0.73	
agree:benefit to user	-0.38	0.23	-1.64	**0.00**	
happening within 2 years (baseline=disagreement)					16751
agreement	-0.72	0.20	-3.7	**0.03**	
data (baseline=specific position)					16872
biometric	0.01	0.24	0.06	0.95	
presence	-2.87	0.35	-8.01	**0.00**	
temperature	-3.66	0.37	-9.66	**0.00**	
video	0.43	0.23	1.82	0.07	
happening today (baseline=disagreement)					17112
agreement	-11.01	349.4	-0.03	0.97	
benefit of scenario (baseline=benefit to company)					18188
benefit to user	-0.46	0.20	-2.30	**0.01**	
purpose unspecified	-1.17	0.27	-4.34	**0.00**	
location (baseline=friend's house)					18569
library	-1.87	0.29	-6.34	**0.02**	
work	-1.96	0.27	-7.34	**0.01**	
house	0.54	0.35	1.52	0.13	
department store	-1.58	0.29	-5.3	**0.00**	
public restroom	-1.23	0.29	-4.17	**0.04**	
retention (baseline=not specified)					18669
week	-0.55	0.11	-4.72	**0.02**	
year	-0.32	0.11	-2.79	**0.00**	
purpose-specific	-0.70	0.12	-5.76	**0.00**	
not deleted	-0.03	0.11	-0.26	0.79	
being shared (baseline=not being shared)					18707
being shared	0.52	0.10	5.41	**0.00**	

Table 7: GLMM Regression Output for the allow-deny model. A positive estimate shows likeliness to deny and a negative estimate shows the likeliness to allow. Factors are ordered by their contribution: the factor with the lowest BIC contributes most to explain participants' desires to allow or deny a data collection.

but this factor was shown to be non-significant in explaining the comfort level. This suggests that being comfortable with a specific data collection instance does not automatically mean that someone would allow it to occur, given the choice.

In the free text answers to the questions about what would make them feel comfortable or uncomfortable with data collection, about 11% of all participants mentioned some type of ability to control collection or use as a requirement for comfort, though our scenarios did not include such a feature. Nevertheless, participants expressed interest in a variety of ways to control their personal information. Within the group that mentioned it, 33% wanted to be granted the ability to delete their data; this would make them feel more comfortable. Another 30% wanted to be asked for consent first, and 27% desired the ability to opt out of the data collection at any time. Multiple participants acknowledged that they would probably not make use of the control options, were they provided.

4.3.2 Predicting Allow/Deny Decisions

Using the significant factors and interactions we determined from the model selection, we trained a machine learning model to predict an individual's decision to allow or deny data collection. The results are shown in Table 6. In this experiment, a prediction is made based on the class (allow or deny) that had the higher probability in the prediction. Averaged over 50 test participants, accuracy ranged from 76% to 80% depending on whether we used most (75%) or all of the other participants' data during training.

Table 6 also describes the results of our simple majority classifier when using all other participant's answers as part of the training data. Similar to when predicting comfort, we use the results of this experiment as an intuitive baseline for understanding how well a classifier does if it simply uses the most prevalent preference in the training data.

The average accuracy of the majority classifier of barely over 50% shows that participants' collective preferences were sufficiently evenly split between wanting to allow and deny data collection in general; hence, a classifier that takes more context into account is necessary for effective prediction. The precision and recall values are 0 because the majority class was always to *deny* data collection, resulting in no true positives ever being predicted, which is clearly not representative of an individual's actual preferences.

Understanding how well we can predict an individual's decision to allow or deny data collection is useful in applications such as where a system pre-populates a privacy control panel with an individual's predicted responses. If an individual changes a pre-populated control (i.e., responding with something different than the system's prediction), the system can update its model with this new "correct" answer. Iteratively refining answers until the system is very confident about a decision will ultimately lead—our results suggest—to the majority of answers specific to an individual being predicted with high confidence.

4.4 Data Collection Notification Preferences

We presented participants with questions asking how often they want to be notified about a data collection with three different frequencies. The frequencies are whether they would want to be notified 1) every time, 2) once in a while, or 3) only the first time the data is collected. They were asked to answer their preferences for all three types of notifications on a five point Likert scale ranging from "Strongly Agree" to "Strongly Disagree."

The best models for describing the three frequencies of notifications revealed that participants' preferences for notification changes based on the factors and levels of factors. The three significant factors that were common between all the models were: data_type, location, and the interaction of these two factors. In these models positive coefficients (estimate) show likeliness of participants' desire to get notification about a data collection.

Factor	Estimate	Std Err	Z-value	p-value	BIC
data:user perceived benefit					13467
baseline=friend's house:not beneficial					
biometrics:not beneficial	0.09	0.21	0.46	0.64	
presence:not beneficial	-0.49	0.24	-2.04	**0.00**	
temperature:not beneficial	-0.38	0.35	-1.1	0.27	
video:not beneficial	0.48	0.22	2.19	**0.00**	
biometrics:purpose unspecified	0.88	0.42	2.12	**0.01**	
presence:purpose unspecified	-0.04	0.48	-0.08	0.93	
temperature:purpose unspecified	-0.71	0.46	-1.55	0.12	
video:purpose unspecified	-0.19	0.47	-0.42	0.67	
data:happening within 2 years					13591
baseline = friend's house:disagree					
video:agree	-0.48	0.34	-1.44	0.15	
biometric:agree	-0.01	0.24	-0.04	0.96	
presence:agree	-0.76	0.33	-2.31	**0.02**	
temperature:agree	-0.11	0.39	-2.28	0.78	
being shared:data (baseline = not being shared:specific position)					13738
being shared:data					13738
baseline = not being shared:specific position					
being shared:presence	0.96	0.22	4.39	**0.00**	
being shared:temperature	-0.27	0.2	-1.32	0.18	
being shared:video	0.73	0.17	4.2	**0.01**	
data (baseline = specific position)					14198
biometric	0.17	0.44	0.39	0.7	
presence	-0.57	0.54	-1.07	0.29	
temperature	-1.66	0.54	-3.07	**0.00**	
video	-0.02	0.52	-0.03	0.98	
happening within 2 years (baseline = disagree)					14697
agree	-0.27	0.19	-1.42	0.15	
user perceived benefit (baseline = beneficial)					14923
not beneficial	0.89	0.16	5.45	**0.00**	
purpose unspecified	0.69	0.35	1.94	**0.04**	
benefit of scenario:location					15281
baseline = benefit to company:friend's house					
benefit to user:department store	-0.01	0.25	-0.02	0.98	
purpose unspecified:department store	0.13	0.28	0.46	0.65	
benefit to user:house	-0.65	0.27	-2.38	**0.01**	
purpose unspecified:library	0.71	0.22	3.18	**0.00**	
benefit to user:library	0.31	0.25	1.28	0.2	
benefit to user:public restroom	0.16	0.25	0.62	0.54	
benefit to user:work	0.29	0.24	1.18	0.23	
benefit of scenario (baseline = benefit to company)					15421
benefit to user	-0.26	0.41	-0.66	0.51	
purpose unspecified	-0.77	0.36	-2.12	**0.00**	
location (baseline = friend's house)					15471
library	-1.11	0.19	-5.58	**0.01**	
work	-1.09	0.19	-5.57	**0.00**	
house	0.79	0.21	3.81	**0.00**	
department store	-0.69	0.20	-3.41	**0.03**	
public restroom	-0.29	0.19	1.48	0.14	
being shared (baseline = not being shared)					15539
being shared	0.17	0.11	1.62	0.11	

Table 8: Generalized Linear Mixed Model Regression output for every-time notification. A positive coefficient (estimate) shows likeliness of participants' desire to get notification about a data collection every time. Factors are ordered by their contribution: the factor with the lowest BIC contributes most to explain participants' preferences about every-time notification.

In the free text answers, 41% of all participants mentioned that being informed would help them feel comfortable, indicated by phrases like "I would want to know..." or "If they would tell me...". Within that group, purpose, a factor heavily dependent on data_type and location, was mentioned by the majority (66%) as something that they would want to be informed about. It was followed by retention (35%), a factor not found in the model. 15% also explicitly requested information on who would be collecting the data (code "collector"). In addition, 13% of this group wanted to be informed about who is accessing the data and 5% want to be informed about steps taken to ensure the security of the collected data. Eight percent of the participants showed some kind of mistrust related to the purpose of data collection described in the scenarios. This was expressed in various ways, from demanding to know "exactly" what was stored and requesting "guarantees" to asking for honesty or expressing general concern about their privacy.

> P928: I like honesty, and with companies being honest and open about why they are sharing data, it makes it a lot easier for me to be comfortable.

More detailed information was also requested about potential risks and how their data was protected against misuse.

4.4.1 Notification Every Time

We measured participants' preferences to get notified about a type of data collection every time it occurred by their answers to the question "I would want my mobile phone to notify me every time this data collection occurs." The factors in the order of their size of effect are shown in Table 8. The most effective factor in explaining participants' desire to be notified every time was the interaction between data_type and user_perceived_benefit, while the factor that had the smallest effect size was shared. Looking at the levels of these factors, it seems that participants were most likely to want to be notified every time when their biometrics were being collected for an unspecified purpose. Also, knowing that the data was being shared had the least effect on participants' desire to be notified every time the data collection occurred.

4.4.2 Notification Once in a While

We measured participants' preferences to being notified only once in a while about a type of data collection by their answers to the question "I would want my mobile phone to notify me every once in a while when this data collection occurs." The results in the order of effect size are shown in Table 9. The model selection algorithm showed that the most effective factor in explaining participants' desire to be notified once in a while was data_type and the least effective factor was the interaction between data_type and location. The coefficients of the levels within these factors show that participants were most likely to want to be notified every once in a while when their biometric was being collected and their desire to get notification every once in a while was least effected by knowing that their presence was being collected while they were at a department store.

4.4.3 Notification the First Time

We measured participants' preferences to being notified only the first time about a type of data collection by their answers to the question, "I would want my mobile phone to notify me only the first time this data collection occurs." Table 10 shows the factors we got from the model selection in order of the effect size. The most effective factor in explaining participants' desire to be notified for the first time was user_perceived_benefit and the factor with the smallest effect size was the interaction between the data_type and location. More specifically, participants were most likely to want to get a notification only the first time if the data collection was not beneficial to them. Also their desire to get notified only for the first time was least effected when their biometric was being collected while they were at a department store.

Factor	Estimate	Std Err	Z-value	p-value	BIC
data (baseline = specific position)					14172
biometric	-0.56	0.16	-3.35	**0.00**	
presence	-0.07	0.24	-0.27	0.78	
temperature	-0.03	0.25	-0.13	0.9	
video	-0.42	0.14	-3.07	**0.01**	
IUIPC					
control	-0.29	0.07	-4.03	**0.00**	14231
location (baseline = friend's house)					14238
library	0.48	0.22	2.21	**0.02**	
work	0.64	0.18	3.63	**0.00**	
house	0.31	0.19	1.63	0.1	
department store	0.29	0.22	1.36	0.18	
public restroom	0.26	0.22	1.19	0.23	
data:location					14243
baseline=specific position:friend's house					
biometric:department store	0.24	0.21	1.14	0.26	
biometric:library	-0.02	0.2	-0.09	0.92	
presence:department store	-0.62	0.29	-2.14	**0.00**	
presence:home	-0.001	0.27	-0.006	0.99	
presence:library	-0.85	0.29	-2.83	**0.00**	
presence:public restroom	-0.67	0.29	-2.29	**0.03**	
presence:work	-0.48	0.25	-1.87	0.61	
temperature:department store	-0.76	0.38	-1.98	**0.00**	
temperature:home	0.52	0.28	1.86	0.62	
temperature:library	-1.34	0.33	-4.06	**0.00**	
temperature:public restroom	-0.86	0.31	-2.87	**0.00**	
temperature:work	-0.87	0.28	-3.12	**0.04**	
video:department store	-0.09	0.19	-0.48	0.62	
video:library	-0.11	0.19	-0.54	0.59	
video:public restroom	-0.30	0.25	-1.20	0.22	

Table 9: Generalized Linear Mixed Model Regression output for once-in-a-while notification. A positive coefficient (estimate) shows likeliness of participants' desire to get notification about a data collection every once in a while. Factors are ordered by their contribution: the factor with the lowest BIC contributes most to explain participants' preferences for once-in-a-while notification.

4.4.4 Summary of Data Collection

At the end of each survey, we asked participants the question "Keeping in mind the 14 scenarios, how often would you be interested in seeing a summary of all such data collection?" Participants could select either every day, every month, every year, or never. Answers varied, with 23% (n = 232) saying they would like a daily summary and 63% (633) selecting a monthly summary. Additionally, 8% (85) would have liked a summary every year and 6% (57) never wanted to receive one.

5. LIMITATIONS

Our study has limitations common to many user studies and to user studies in the area of privacy. Although the demographic attributes of the participant group are, except for the reported income, close to the US average, Mechanical Turk workers do not reflect the general population. Prior research has shown that Mechanical Turk workers are more privacy-sensitive than the general population [16]. It has also been has shown that self reports about privacy preferences often differ from actual behavior. This is referred to as the "privacy paradox" [10, 1]. Our study may be susceptible to this bias because the scenarios were abstract and participants were asked to imagine themselves in situations they may not have encountered. In addition,

Factor	Estimate	Std Err	Z-value	p-value	BIC
user perceived benefit (baseline=beneficial)					14487
not beneficial	-0.47	0.07	-7.09	**0.01**	
purpose unspecified	-0.32	0.05	-6.08	**0.00**	
location (baseline=friend's house)					14567
library	0.74	0.22	3.37	**0.02**	
work	0.86	0.18	4.76	**0.00**	
house	0.08	0.19	0.41	0.68	
department store	0.75	0.22	3.36	**0.03**	
public restroom	0.61	0.22	2.81	**0.00**	
data (baseline=specific position)					14587
biometric	0.17	0.17	1.02	0.31	
presence	0.78	0.24	3.24	**0.00**	
temperature	0.81	0.25	3.30	**0.00**	
video	0.00	0.13	-0.02	0.99	
data:location					14617
baseline = specific position:friend's house					
biometric:department store	-0.58	0.21	-2.79	**0.00**	
biometric:library	-0.30	0.2	-1.51	0.13	
presence:department store	-1.05	0.29	-3.66	**0.00**	
presence:home	-0.23	0.27	-0.83	0.41	
presence:library	-1.19	0.29	-4.02	**0.02**	
presence:public restroom	-1.19	0.29	-4.13	**0.00**	
presence:work	-0.48	0.25	-1.86	0.06	
temperature:department store	-1.61	0.38	-4.26	**0.00**	
temperature:home	0.23	0.28	0.82	0.41	
temperature:library	-1.35	0.32	-4.18	**0.00**	
temperature:public restroom	-1.09	0.31	-3.58	**0.00**	
temperature:work	-1.17	0.28	-4.19	**0.01**	
video:department store	-0.16	0.19	-0.85	0.39	
video:library	-0.17	0.19	-0.89	0.37	
video:public restroom	-0.54	0.25	-1.20	0.22	

Table 10: Generalized Linear Mixed Model Regression output for first-time-only notification. A positive coefficient (estimate) shows likeliness of participants' desire to get notification about a data collection only the first time. Factors are ordered by their contribution: the factor with the lowest BIC contributes most to explain participants' preferences for first-time-only notification.

some of the scenarios in our study were designed to be realistic based on common data collection and use practices that are happening today, while others were designed to be more forward-looking. We decided to have some less-realistic scenarios because we hypothesized that there is a relation between participants' comfort level about each vignette and their perception of how realistic it is. Nevertheless, participants may have been asked about situations which they are not typically put in, influencing their decisions.

Despite these limitations, presenting a large variety of scenarios to participants allowed us to explore situations that do not currently happen but may be similar to situations that will happen in the future. Since the Internet of Things is still an emerging field, it is not possible to describe situations that are realistic to all participants who may never have had an IoT device or never have faced a situation in which an IoT sensor is collecting data.

6. DISCUSSION

Our results demonstrate varied privacy concerns, both across IoT scenarios and across participants. Our results also indicate that participants are more comfortable about data collection when classical privacy and data protection rules, such as the Fair Information Practices, are applied and individuals are given an explanation about why their data is being collected. However, other results underline the need for technology to support the awareness of data collection and that can meet the different desires for being notified.

6.1 Privacy Preferences Are Complex

How individuals feel about different data collection scenarios depends on various things. Individual preference play as much a role as social norms and expectations.

On one hand, our analyses show that participants are largely in agreement on a number of practices where social norms are in place that define what is acceptable and what is not. For example, participants expressed more comfort with data collection in public spaces, but rejected scenarios that described video cameras used in private rooms and shared with law enforcement. This is likely related to a long, western tradition of public/private dichotomy. However, this dichotomy is challenged by smart-home technology with centralized, cloud-based services that do not follow expectation of "what happens at home stays at home." For example, Samsung received criticism for advising the public not to have private conversations in front of their smart TV [14] as it uses a third party speech-to-text service for voice commands. Smart-home device manufacturers should be aware and respectful of individuals' mental models of data collection within the home and do their best to communicate practices that may be surprising to their customers.

On the other hand, we saw a large number of scenarios in which there was no clear indication of what is generally acceptable. For example, participants showed a high variance in the level of comfort with respect to the collection and storage of movement patterns at their workplace for the purpose of optimizing heating and cooling. Social norms have yet to emerge with respect to technology that has just recently become available. However, scenarios like these also reflect how individual preferences might differ in the long run. Individuals have to weigh their potential loss of privacy, due to camera surveillance against the benefit of reduced energy consumption. The complexity of this individual decision process is also reflected by the fact that our models describing the comfort level and the choice to allow or deny a data collection do not completely overlap. Here individual concerns about what might happen to the data, in combination with personal experience (e.g., how much one trusts her employer), play a role in determining whether or not one feels comfortable with the data collection and will allow it.

6.2 Addressing Privacy Concerns

Both the qualitative and quantitative data show that participants prefer anonymous data collection. Temperature and presence sensors produce data that are not immediately identifying and participants consistently expressed higher comfort with these scenarios. This finding was further reinforced by our free-text results, as anonymous data was the second most mentioned preference for data collection. This is further confirmed through interviews done in a previous study [7]. The relatively high discomfort with data inference, combined with high comfort regarding collection of anonymous data indicates that people may be generally unaware that with the Internet of Things it will be easier to re-identify individuals from otherwise anonymous data. In light of our findings, it is likely that this is something that would cause discomfort. This gap in understanding should be kept in mind when providing privacy information for IoT data collection.

We found that participants favor short retention times and are more comfortable when data is deleted after its purpose is met, or not kept longer than a week. Insights from the free-text responses indicate that this is related to an increased awareness of data breaches, the fear of misuse of data, and concerns regarding bad data security practices at companies. As previous research has shown, a growing number of people have already experienced misuse of their data [34]. With the growing number of IoT devices, the probability of data breaches further increases, resulting in higher concern and less trust in the technology. To address these types of concerns, IoT device manufacturers should take precautions, both technical and administrative, to protect their customers' data and communicate these practices to the public.

6.3 Towards Awareness and Control

Approaches for eliciting consent or providing information are less likely to work in the IoT setting. For example, a classic privacy policy cannot be shown on many types of IoT devices, such as a smart watch. Still, people demand information about the entity collecting data, the purpose of the collection, the benefit they receive from it, and the retention period of the collected data.

In open-ended responses, participants explicitly asked for transparency in data collection and its handling. Discomfort increases when data is shared with third parties or used to infer additional information. Participants want to be informed not only about the purpose of data collection and the handling of data, but also possible security risks associated. This finding is also confirmed by previous work which found through interviews that transparency about the data collected and the purpose of the collection influence comfort levels for data collection by IoT devices [7].

Additionally, our results show that how often and about what participants want to be informed is greatly dependent on individual comfort levels. But information requests also heavily depend on whether or not individuals think a use of their data is beneficial to them or serves a greater good. To answer this question even semi-automatically requires more specific and neutral information about the purpose of a data collection. We also saw that two thirds of participants would appreciate a monthly summary about what data has been collected about them (see section 4.4).

To develop technical support for this is a major challenge in a fractured IoT landscape that still lacks standardization. One option to streamline these efforts, at least on a smaller scale like in smart homes, would be to build upon the Manufacture Usage Description Specification [11] to include information on purposes of data collection and simplify the aggregation of information about data collection.

Our analysis suggests that many people want to retain control of their personal data. Future IoT services should take this into consideration when designing privacy notices instead of creating more "one-size fits all" policies.

More specifically, we suggest the adoption of the idea of personalized privacy assistants (PPA) already used in the context of mobile apps [25]. A PPA may be a tool or agent running on behalf of each individual that can proactively predict their decision to allow or deny data collection, relieving the individual of making decisions when they can be predicted with high accuracy. This predictive model could be used to, i.e., pre-populate a privacy control panel with individuals' preferences. In a deployed system, we could use a form of online machine learning to continue to update the model to a specific individual's preferences. Our predictive model 4.3 showed that with a few data points per individual (three), we could predict the rest of their eleven answers with an average accuracy of 88%. In a deployed system, we expect the model would have more specific data points about individuals on which to base predictions, which would be even more accurate.

7. CONCLUSIONS

In this paper we reported on a large-scale vignette study on privacy concerns related to the Internet of Things. We asked 1,007 participants to rate realistic scenarios about data collection occurring in multiple contexts. Our results enhance the findings of previous, mostly qualitative research with statistical evidence that identifies specific factors that impact individuals' privacy concerns. Among these factors are the type of data that is collected, retention time, third-party sharing, perceived benefit, and the location at which an IoT device collects data. The statistical results are confirmed by analyses of the free-text responses, which emphasize concerns regarding the collection of biometric data as well as data collection occurring in private spaces.

Based on our findings, we made recommendations for designing IoT services and applications. People favor data collection in which they cannot be identified immediately. They also do not want inferences to be made from otherwise anonymous data. We found that participants want to be informed about various details of data collection, such as what the data is used for and how long it will be stored.

8. ACKNOWLEDGEMENTS

This research has been supported in part by DARPA and the Air Force Research Laboratory under agreement number FA8750-15-2-0277 and by the National Science Foundation under grant SBE-1513957. The US Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright notation thereon. Additional support has also been provided by Google. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, the Air Force Research Laboratory, the NSF, Google, or the US Government.

9. REFERENCES

[1] Alessandro Acquisti and Ralph Gross. 2006. Imagined communities: Awareness, information sharing, and privacy on the Facebook. In *Proc. PETS*.

[2] Ivor D. Addo, Sheikh Iqbal Ahamed, Stephen S. Yau, and Arun Balaji Buduru. 2014. A Reference Architecture for Improving Security and Privacy in Internet of Things Applications. In *IEEE Third International Conference on Mobile Services*. 108–115. DOI: http://dx.doi.org/10.1109/MobServ.2014.24

[3] Christiane Atzmüller and Peter M. Steiner. 2010. Experimental vignette studies in survey research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 6, 3 (2010), 128–138. DOI: http://dx.doi.org/10.1027/1614-2241/a000014

[4] Debjanee Barua, Judy Kay, and Cécile Paris. 2013. Viewing and Controlling Personal Sensor Data: What Do Users Want? In *Persuasive Technology*, Shlomo Berkovsky and Jill Freyne (Eds.). Number 7822 in Lecture Notes in Computer Science. Springer Berlin Heidelberg, 15–26. http://link.springer.com/chapter/10.1007/978-3-642-37157-8_4 DOI: 10.1007/978-3-642-37157-8_4.

[5] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. DOI: http://dx.doi.org/10.18637/jss.v067.i01

[6] Pankaj Bhaskar and Sheikh Iqbal Ahamed. 2007. Privacy in Pervasive Computing and Open Issues. In *Proceedings of the The Second International Conference on Availability, Reliability and Security, ARES 2007, The International Dependability Conference - Bridging Theory and Practice*. 147–154. DOI: http://dx.doi.org/10.1109/ARES.2007.115

[7] Igor Bilogrevic and Martin Ortlieb. 2016. "If You Put All The Pieces Together...": Attitudes Towards Data Combination and Sharing Across Services and Companies. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. DOI: http://dx.doi.org/10.1145/2858036.2858432

[8] Richard Chow, Serge Egelman, Raghudeep Kannavara, Hosub Lee, Suyash Misra, and Edward Wang. 2015. HCI in Business: A collaboration with academia in IoT privacy. In *International Conference on HCI in Business*. Springer, 679–687.

[9] The Federal Trade Commission. 2015. *Internet of Things: Privacy & Security in a Connected World*. Technical Report. Federal Trade Commission. Accessed Mar. 2017.

[10] Catherine Dwyer, Starr Roxanne Hiltz, and Katia Passerini. 2007. Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace. In *Proc. AMCIS*.

[11] E. Lear, R. Droms, and D. Romascanu. 2017. *Manufacturer Usage Description Specification*. Internet-Draft draft-ietf-opsawg-mud-04. IETF Network Working Group. https://datatracker.ietf.org/doc/draft-ietf-opsawg-mud/?include_text=1

[12] Serge Egelman, Raghudeep Kannavara, and Richard Chow. 2015. Is this thing on?: Crowdsourcing privacy indicators for ubiquitous sensing platforms. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1669–1678.

[13] Janet Finch. 1987. The vignette technique in survey research. *Sociology* (1987), 105–114.

[14] David Goldman. 2015. Your Samsung TV is eavesdropping on your private conversations. (Feb. 2015). http://money.cnn.com/2015/02/09/technology/security/samsung-smart-tv-privacy/index.html

[15] Joseph B Kadane and Nicole A Lazar. 2004. Methods and criteria for model selection. *Journal of the American statistical Association* 99, 465 (2004), 279–290.

[16] Ruogu Kang, Stephanie Brown, Laura Dabbish, and Sara B Kiesler. 2014. Privacy Attitudes of Mechanical Turk Workers and the US Public. In *SOUPS*. 37–49.

[17] Predrag Klasnja, Sunny Consolvo, Tanzeem Choudhury, Richard Beckwith, and Jeffrey Hightower. 2009. Exploring Privacy Concerns about Personal Sensing. In *Pervasive Computing*, Hideyuki Tokuda, Michael Beigl, Adrian Friday, A. J. Bernheim Brush, and Yoshito Tobe (Eds.). Number 5538 in Lecture Notes in Computer Science. Springer Berlin Heidelberg, 176–183. http://link.springer.com/chapter/10.1007/978-3-642-01516-8_13 DOI: 10.1007/978-3-642-01516-8_13.

[18] Scott Lederer, Jason I Hong, Anind K Dey, and James A Landay. 2004. Personal privacy through understanding and action: five pitfalls for designers. *Personal and Ubiquitous Computing* 8, 6 (2004), 440–454.

[19] Scott Lederer, Jennifer Mankoff, and Anind K. Dey. 2003. Who wants to know what when? Privacy preference determinants in ubiquitous computing. In *Extended abstracts of the 2003 Conference on Human Factors in Computing Systems, CHI 2003, Ft. Lauderdale, Florida, USA, April 5-10, 2003*. 724–725. DOI: http://dx.doi.org/10.1145/765891.765952

[20] Hosub Lee and Alfred Kobsa. 2016. Understanding User Privacy in Internet of Things Environments. *Internet of Things (WF-IoT)* (2016).

[21] Hosub Lee and Alfred Kobsa. 2017. Privacy Preference Modeling and Prediction in a Simulated Campuswide IoT Environment. In *Proceedings of the 15th IEEE Conference on Pervasive Computing and Communications*. IEEE.

[22] Pedro Giovanni Leon, Blase Ur, Yang Wang, Manya Sleeper, Rebecca Balebako, Richard Shay, Lujo Bauer, Mihai Christodorescu, and Lorrie Faith Cranor. 2013. What matters to users?: factors that affect users' willingness to share information with online advertisers. In *Proceedings of the ninth symposium on usable privacy and security*. ACM, 7.

[23] Jialiu Lin, Shahriyar Amini, Jason I Hong, Norman Sadeh, Janne Lindqvist, and Joy Zhang. 2012. Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 501–510.

[24] Jialiu Lin, Bin Liu, Norman Sadeh, and Jason I Hong. 2014. Modeling users'ĂŹ mobile app privacy preferences: Restoring usability in a sea of permission settings. In *Symposium on Usable Privacy and Security (SOUPS)*, Vol. 40.

[25] Bin Liu, Mads Schaarup Andersen, Florian Schaub, Hazim Almuhimedi, Shikun Aerin Zhang, Norman Sadeh, Yuvraj Agarwal, and Alessandro Acquisti. 2016. Follow My Recommendations: A Personalized Assistant for Mobile App Permissions. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*.

[26] Bin Liu, Jialiu Lin, and Norman Sadeh. 2014. Reconciling mobile app privacy and usability on smartphones: Could user privacy profiles help?. In *Proceedings of the 23rd international conference on World wide web*. ACM, 201–212.

[27] Naresh K Malhotra, Sung S Kim, and James Agarwal. 2004. Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information systems research* 15, 4 (2004), 336–355.

[28] Kirsten E Martin. 2012. Diminished or just different? A factorial vignette study of privacy as a social contract. *Journal of Business Ethics* 111, 4 (2012), 519–539.

[29] Kirsten E Martin and Helen Nissenbaum. 2016. Measuring Privacy: An Empirical Test Using Context To Expose Confounding Variables. *Columbia Science and Technology Law Review* 18 (2016), 176–218.

[30] Sharad Mehrotra, Alfred Kobsa, Nalini Venkatasubramanian, and Siva Raj Rajagopalan. 2016. TIPPERS: A privacy cognizant IoT environment. In *Pervasive Computing and Communication Workshops (PerCom Workshops), 2016 IEEE International Conference on*. IEEE, 1–6.

[31] Helen Nissenbaum. 2009. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[33] Charith Perera, Rajiv Ranjan, Lizhe Wang, Samee Ullah Khan, and Albert Y. Zomaya. 2015. Big Data Privacy in the Internet of Things Era. *IT Professional* 17, 3 (2015), 32–39. DOI:http://dx.doi.org/10.1109/MITP.2015.34

[34] Lee Rainie, Sara Kiesler, Ruogu Kang, and Mary Madden. 2013. Anonymity, Privacy, and Security Online. (Sept. 2013). http://www.pewinternet.org/2013/09/05/anonymity-privacy-and-security-online/

[35] Beate Rössler. 2005. *The value of privacy* (english ed ed.). Polity, Cambridge, UK ; Malden, MA.

[36] Norman Sadeh, Jason Hong, Lorrie Cranor, Ian Fette, Patrick Kelley, Madhu Prabaker, and Jinghai Rao. 2009. Understanding and capturing people's privacy policies in a mobile social networking application. *Personal and Ubiquitous Computing* 13, 6 (2009), 401–412.

[37] Janice Y Tsai, Patrick Kelley, Paul Drielsma, Lorrie Faith Cranor, Jason Hong, and Norman Sadeh. 2009. Who's viewed you?: the impact of feedback in a mobile location-sharing application. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2003–2012.

[38] Arijit Ukil, Soma Bandyopadhyay, and Arpan Pal. 2015. Privacy for IoT: Involuntary privacy enablement for smart energy systems. In *2015 IEEE International Conference on Communications, ICC 2015, London, United Kingdom, June 8-12, 2015*. 536–541. DOI: http://dx.doi.org/10.1109/ICC.2015.7248377

APPENDIX

Appendix material is formatted differently than what appeared in the survey seen by participants.

A. SAMPLE SURVEY SCENARIO

You are at a **friend's house**. All rooms have **presence sensors that are used to determine when to switch on and off the lights to reduce costs and save energy**. You are **not told how long the data will be kept.**

Q1. This use of my data would be beneficial to me. (Answered on a five point Likert scale from "Strongly Agree" to "Strongly Disagree")

Q2. I think scenarios like this happen today. (Answered on a five point Likert scale from "Strongly Agree" to "Strongly Disagree")

Q3. (If "disagree" or "strongly disagree" for Q2) I think scenarios like this will happen within 2 years. (Answered on a five point Likert scale from "Strongly Agree" to "Strongly Disagree")

Q4. (If "disagree" or "strongly disagree" for Q3) I think scenarios like this will happen within 10 years. (Answered on a five point Likert scale from "Strongly Agree" to "Strongly Disagree")

Q5a. How would you feel about the data collection in the situation described above if you were not told with whom the data would be shared, how long it would be kept or how long it would be used for? (Answered on a five point Likert scale from "Very Comfortable" to "Very Uncomfortable")

Q5b. How would you feel about the data collection in the situation described above if you were given no additional information about the scenario? (Answered on a five point Likert scale from "Very Comfortable" to "Very Uncomfortable")

Q6a. I would want my mobile phone to notify me every time this data collection occurs. (Answered on a five point Likert scale from "Strongly Agree" to "Strongly Disagree")

Q6b. I would want my mobile phone to notify me only the first time this data collection occurs. (Answered on a five point Likert scale from "Strongly Agree" to "Strongly Disagree")

Q6c. I would want my mobile phone to notify me every once in a while when this data collection occurs. (Answered on a five point Likert scale from "Strongly Agree" to "Strongly Disagree")

Q7. If you had the choice, would you allow or deny this data collection? (Choices: Allow, Deny)

B. SUMMARY QUESTIONS

Q1. Keeping in mind the 14 scenarios, how often would you be interested in seeing a summary of all such data collection? (Choices: Every day, Every month, Every year, Never)

Q2. Keeping in mind the 14 scenarios, what would make you comfortable with sharing data in such situations?

Q3. Keeping in mind the 14 scenarios, what would make you uncomfortable with sharing data in such situations?

C. IUIPC QUESTIONS

Participants answered the following questions on a seven point Likert scale from "Strongly Aagree" to "Strongly Disagree"

1. 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.
2. Consumer control of personal information lies at the heart of consumer privacy.
3. I believe that online privacy is invaded when control is lost or unwillingly reduced as a result of a marketing transaction.
4. Companies seeking information online should disclose the way the data are collected, processed, and used.
5. 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.
6. It usually bothers me when online companies ask me for personal information.
7. When online companies ask me for personal information, I sometimes think twice before providing it.
8. It bothers me to give personal information to so many online companies.
9. I'm concerned that online companies are collecting too much personal information about me.

D. DEMOGRAPHIC QUESTIONS

Q1. How old are you?

Q2. What is your gender? (Choices: Female, Male, Other, Prefer not to answer)

Q3. What is the highest degree you have earned? (Choices: No high school degree, High school degree, College degree, Professional degree (masters/PhD), Associates degree, Medical degree, Prefer not to answer)

Q4. What is your income range? (Choices: Less than $15,000/ year, $15,000/ year - $24,999/year, $25,000/ year - $34,999/ year, $35,000/ year - $49,999/ year, $50,000/ year - $74,999/ year, $75,000/ year - $99,999/ year, $100,000/ year - $149,999/year, $150,000/year - $199,999/ year, $200,000/ year and above, Prefer not to answer)