"Millions of people are watching you": Understanding the Digital-Safety Needs and Practices of Creators

Patrawat Samermit, Anna Turner, Patrick Gage Kelley, Tara Matthews, Vanessia Wu, Sunny Consolvo, and Kurt Thomas, *Google*

https://www.usenix.org/conference/usenixsecurity23/presentation/samermit

This paper is included in the Proceedings of the 32nd USENIX Security Symposium.

August 9–11, 2023 • Anaheim, CA, USA

978-1-939133-37-3

Open access to the Proceedings of the 32nd USENIX Security Symposium is sponsored by USENIX.

"Millions of people are watching you": Understanding the Digital-Safety Needs and Practices of Creators

Patrawat Samermit Anna Turner Patrick Gage Kelley Tara Matthews
Vanessia Wu Sunny Consolvo Kurt Thomas
{patrawat, annaturn, patrickgage, taramatthews, vanessia, sconsolvo, kurtthomas}@google.com
Google

Abstract

Online content creators—who create and share their content on platforms such as Instagram, TikTok, Twitch, and YouTube—are uniquely at-risk of increased digital-safety threats due to their public prominence, the diverse social norms of wide-ranging audiences, and their access to audience members as a valuable resource. We interviewed 23 creators to understand their digital-safety experiences. This includes the security, privacy, and abuse threats they have experienced across multiple platforms and how the threats have changed over time. We also examined the protective practices they have employed to stay safer, including tensions in how they adopt the practices. We found that creators have diverse threat models that take into consideration their emotional, physical, relational, and financial safety. Most adopted protections—including distancing from technology, moderating their communities, and seeking external or social support—only after experiencing a serious safety incident. Lessons from their experiences help us better prepare and protect creators and ensure a diversity of voices are present online.

1 Introduction

Online content creators (referred to simply as "creators" throughout this paper) are people who create and share their content with online audiences—from large to small—on platforms such as Instagram, TikTok, Twitch, and YouTube. For some, creating is a full-time job supported by ads, merchandise, subscriptions, and brand deals [4]. For others, it's a creative outlet or a path towards professional independence. A creator's online presence spans a dizzying number of mediums and platforms. Short- and long-form videos, photos, comments, reactions, text posts, community pages, and livestreams all work towards a creator's goal of elevating their voice and establishing their brand.

A creator's heightened visibility and fingertips-away availability to their audience places them at higher risk of digital-safety threats. For example, Dream, a popular Minecraft player with 31 million YouTube subscribers as of December 2022, was doxxed when his residential address was posted by an attacker who correlated a photo Dream shared of his kitchen with a real estate listing photo found on Zillow [9]. Black and LGBTQ+ streamers on Twitch recently experienced coordinated harassment campaigns where thousands of automated bots posted toxic messages to their streams [13]. And hack-for-hire criminal groups have targeted creators to disseminate cryptocurrency scams to their audiences and to siphon their ad revenue [22]. While these encounters are infrequent, they represent a wide-range of threats that creators could—but hopefully never will—experience.

As barriers to creating lower, more people are interacting with broad online audiences. To support this shift, the security and HCI communities can benefit from greater understanding of how creators—from those who are well-established to those early in their journeys—think about and protect their digital safety.[1] We interviewed 23 creators who represent a diversity of voices with audience sizes ranging from 5,000 to over 750,000 to understand:

RQ1: Risks. What contributes to creators' unique, pervasive, and/or severe digital-safety risks? How are risks impacted by creator's identities, content types, and audience expectations? How do their experiences relate to those of other at-risk populations?

RQ2: Threat models. What are creators' top perceived or experienced digital-safety threats? What attackers and potential harms are they concerned about?

RQ3: Protective practices. What protective practices do creators adopt? How and when did they learn about and adopt the practices? What challenges did they face enacting the practices?

Creators in our study emphasized the positives they experienced, such as building a community with their audiences, having an outlet for their creativity that they could share with the

[1] We use the term, *digital safety*, to encapsulate security, privacy, abuse, or other online safety risks.

world, and developing a business from something they love. However, they also reported facing a cross-platform threat landscape where attacks could come from anyone at any time, including anonymous online attackers, established audience members, other creators, family and friends, and scammers. Confirming findings from our prior survey [27], we found that creators face a host of attacks including toxic content, content leakage, stalking, and more. Expanding on this prior work, we synthesize creators' unique set of risk factors and explore in-depth the emotional, physical, relational, and financial harms creators experienced, including the trade-offs they made in an attempt to keep themselves—and others—safe. For example, their exposure and accessibility as a creator typically helped them financially, but often came at the cost of emotional, and sometimes physical, safety. Certain kinds of harm—like relational damage or loss of privacy (e.g., due to leaked information)—were hard (or impossible) to reverse. Maintaining community safety required moderation that exposed creators to further emotional harms. These trade-offs were not readily resolved with existing platform tools. Further, the harms involved in these trade-offs were potentially elevated for creators with marginalized identities or characteristics, whose risk of attack was intersectional.

Most creators did not enact sufficient protections until *after* they experienced harm. Their protections focused on avoiding certain features, limiting sharing, or even leaving platforms to maintain their physical and emotional safety; employing moderation to protect their relational and emotional safety; and seeking social and external support for a range of issues not well supported by current tooling. We distill our results into recommendations that aim to help overcome digital-safety barriers that creators experience, and in doing so, chart a path towards elevating the digital safety of creators to ensure a diversity of voices are present online.

2 Related Work

We summarize prior work related to digital-safety threats faced by creators, the digital-safety experiences of users more broadly who are at higher risk, and frameworks we employ.

Understanding threats. While the media has covered specific attacks on creators, limited scholarly work has deeply explored creators' digital-safety experiences. Previous work has examined how creators manage interactions with their audiences [30], cope with negativity [28], can use collaborative filtering to help manage toxicity by sharing keyword filters [14], experience unique risks when engaging in online sex work [12], and face gender-based risks associated with higher visibility online [8, 23].

Most related to this work is our 2021 survey of 135 creators [27] that quantified the hate and harassment attacks creators experienced and coping practices used in response to the attacks. We found that hate and harassment was incredibly common (experienced by 95% of respondents across all platforms they participated on), with over half of respondents being moderately, or more, concerned about future attacks.

We build on this work qualitatively, providing more context on the factors that increase creators' risk of being attacked and in-depth stories and incidents that add to the community's understanding of how attacks, harms, and practices are interconnected. We also expand beyond creators' experiences with hate and harassment to digital-safety harms more broadly, the protective practices they employed, and barriers they encountered in protecting themselves. Our study provides new insights and rich nuance about the digital-safety experiences of creators and their short- and long-term needs.

Exploring occupations. Though there is limited digital-safety research on creators, interest in the digital-safety experiences of at-risk users is rising, including within the security community. This has yielded many studies on marginalized groups and other populations who face heightened digital-safety threats. Most related to our work are studies focused on occupations or activities that increase risk, including activists [7, 25], journalists [18, 19], sex workers [2, 17, 24], and political campaigners [6]. Stories from the creators in our study add to an increasing understanding of how one's occupation can impact one's digital safety. Where feasible, we contrast our findings with these other populations to understand the unique intersection of risk factors for creators.

Employing frameworks. Our work builds on frameworks for reasoning about how people experience and cope with digital-safety threats. We demonstrate that creators are *at-risk users*, defined by Warford et al. [29] as people "who experience risk factors that augment or amplify their chances of being digitally attacked and/or suffering disproportionate harm." Using Warford et al.'s framework, we examine how *prominence*, *social norms*, *marginalization*, and *access to a sensitive resource* manifest for creators as risk factors (i.e., factors that augment or amplify their digital-safety risk) [29]. To explore attacks creators face, we rely on the hate and harassment attack categories—e.g., toxic content, content leakage, and false reporting—from Thomas et al. [26]. To understand harms experienced by creators, we rely on a framework from Scheuerman et al. [20]—which categorizes physical, emotional, financial, and relational harms—and expand their definition of relational harm to include *community-based harms*.

3 Methodology

We interviewed 23 creators to explore their digital-safety experiences, concerns, and protective practices in depth. We describe our study's participants, data collection, analysis approach, ethical practices, and limitations.

3.1 Participants & recruiting

We recruited from a large pool of creators who had opted-in to a program for those interested in participating in research. This program was managed by YouTube [31], which invested resources to ensure a diverse set of potential participants in terms of gender identity, race/ethnicity, and identification within the LGBTQ+ community.

Participation in our study was restricted to US-based creators who were 18+ years old. We purposefully recruited participants who created on two or more platforms, and who represented a wide range of content verticals and audience sizes. Though all participants created on YouTube, they reported using a median of 5 platforms each. The platforms they created on included Amazon Live, Discord, Facebook, Instagram, Reddit, Snapchat, TikTok, Twitch, Twitter, and YouTube, with participants relying on platforms like Patreon and LinkedIn for managing their audience and monetizing their brand. Their content verticals included beauty, education, entertainment, fitness, gaming, lifestyle, news, and vlogging. In terms of their audience sizes (i.e., their number of followers or subscribers), six participants had between 5,000-24,999 followers on their largest platform; three had 25,000-99,999; 10 had 100,000-749,999; and four had more than 750,000 followers. Participants had been creating—full or part-time—over the last 2.5-15 years.

To protect participants' privacy, we did not collect personal demographic information, so we cannot provide additional details.[2] Participants received $100 USD as a thank you gift.

3.2 Data collection

We conducted one-on-one semi-structured interviews with the 23 creators from November 2021-January 2022. Interview sessions, which were recorded, were virtual and ran 82 minutes on average. Interviews were comprised of four sections:

(1) Career. We asked about their history as a creator, including the platforms they used and their dependence on monetization.

(2) Safety concerns. We delved into their top digital-safety concerns, including the origin of their concerns (e.g., personal experience, observing the experiences of others, etc.).

(3) Protective practices. We asked about what they did to stay safe, and from where or whom they learned those practices (e.g., another creator, a security expert, the internet, etc.).

(4) Advice & resources. We discussed what they would have liked to have known when starting their creating journey, and whether they knew of any helpful digital-safety resources.

[2] Some quotes in this paper include demographic information. Such details are included only when participants chose to share.

3.3 Analysis approach

We used a thematic analysis [3] to analyze the interview data. Thematic analysis was appropriate because our goal was to produce a set of patterns coupled with rich detail that would be useful for building an in-depth understanding of creators' digital-safety experiences. After familiarizing ourselves with the collected data, we developed an initial codebook of inductive and deductive codes. To build on existing knowledge about at-risk users, we selected deductive codes from previous frameworks including hate and harassment attacks [26], the types of harms that result from abusive content [20], and common protective practices adopted by at-risk users [29]. We expanded these to include inductive codes related to factors that amplified risk for creators, further nuance on attacks and harms, protective practices specific to creating (e.g., moderation), where creators learned about protective practices, and critical moments where creators formed or refined their threat models. Our inductive codes started as high-level categories that we developed throughout coding, ultimately allowing us to identify novel, data-driven themes.

Our data consisted of nearly 32 hours of video recordings. Transcripts were generated from the recordings automatically. Six research team members used the codebook to code the recordings and transcript data. Changes to the codebook were discussed and agreed upon as coding progressed. As we coded, we wrote summaries and memos to synthesize potential themes. Once coding was complete, we searched for themes by collating and reviewing coded data and memos. We outlined and discussed key themes, ultimately selecting those that were both pervasively reported by participants and important for mitigating digital-safety risks to creators. We performed credibility checks by discussing our findings with domain experts who work extensively with creators.

3.4 Ethics

Our study plan was reviewed by experts at our institution in domains including ethics, human subjects research, policy, legal, security, privacy, and anti-abuse. Our institution does not have an IRB, though we adhere to similarly strict standards. Before conducting our interviews, we received informed consent from each participant. At the onset of each interview, we reminded participants to only share what they were comfortable sharing, and that they could stop the interview or recording at any time—they would receive their thank you gift regardless.

To protect participants' privacy, recruitment and distribution of the thank you gift were handled by the participant recruitment program coordinators (who were not on the research team). Access to recordings and transcripts of interview sessions were restricted to the research team (and institutional administrators). When reporting participant quotes and stories, we omit unique details, phrases, or words to minimize the risk of participants being identified. When a quote relates to a

participant's marginalized identity, we omit the participant's pseudonymous code to further protect their identity.

3.5 Limitations

In addition to standard limitations of self-reported data (e.g., recall and observer biases), participants may have minimized, omitted, or failed to recall digital-safety concerns and protective practices that were not top-of-mind or that they were uncomfortable sharing. Our participants were US-based creators who were 18+ years old, so our findings may not generalize to other ages or cultural contexts. All participants created on YouTube and were recruited via a YouTube research program, though they also all created on multiple platforms (with a median of 5 platforms). Given the popularity of the creators we spoke to (roughly half of whom had 100,000+ viewers or followers), their experiences likely skew towards more frequent threats compared to a random sample of creators, as identified in our prior work [27]. Nevertheless, our findings highlight wide-ranging threats that creators can experience across platforms, underscoring their at-risk context.

Our findings include intersectional risk factors that contribute to participants' digital-safety concerns and experiences, but our method does not allow us to disambiguate the root cause of specific safety issues, or whether one risk factor contributes more than another to reducing a creator's safety. Future work comparing populations with a subset of, or disjoint risk factors may help in pinpointing how specific risk factors correlate with heightened digital-safety needs. Such studies would also need to account for the roles of platform affordances and monetization as they apply to digital safety.

4 Risk Factors of Creating

Creators in our study highly valued their status as public figures, emphasizing benefits such as the sense of community that came from creating. But their accessibility to the public exposed these creators to potentially pervasive and severe risks, aligning with the *prominence*, *social norms*, *marginalization*, and *access to a sensitive resource* risk factors associated with at-risk populations from Warford et al. [29]. Risk factors represent unique circumstances that augment or amplify a person's chances of being digitally attacked and/or suffering disproportionate harm, thus putting them at-risk. We first explore participants' risk factors, underscoring the unique set of intersectional risks that creators in our study experienced as a result of their occupation or hobby, and why. Then we describe why participants created despite these risks.

4.1 Prominence

Participants' public prominence exposed them to broad online audiences, which placed them at higher risk of digital-safety attacks. While prominence is a risk factor also associated with politicians [6] and journalists [18, 19] (among others), we identify themes unique to creators.

Audiences seemed to expect frequent & authentic interactions with creators. Creators in our study perceived themselves as different from "traditional" celebrities (like actors, musicians, or other public figures), because they felt their audiences expected frequent, authentic interactions and access to their thoughts and experiences. They attributed this expectation, in part, to platform features that made them highly accessible to community members (e.g., via comments, likes, shout-outs, etc.).

> "[Creators] are easier to target than [traditional celebrities]... It's more accessible to make a comment on something right under a creator's content. For a movie, you can't write something on [a streaming service] under the movie. The movie is just there for you to watch and consume. It's not there for you to necessarily provide your input." – C14

Creators in our study expressed that—unlike traditional celebrities—their regular engagement with their audience felt *necessary* to maintain their online presence and grow their audience. This engagement was deeply entangled with a creator's financial success, and thus at odds with efforts to scale back touch-points or interactions to ensure safety or autonomy. C17 shared a poignant example of this tension:

> "If somebody gives me $10, they feel like... I am indebted to them. Once you entangle money with it, it becomes a lot more difficult to tell somebody 'no.' A friend of mine who's a streamer... [They're] growing really rapidly. [They've] got people throwing $400 at [them], on a single stream and [they're] having that same problem of, 'I don't know how to react to these people. I don't know how to handle this. This guy just gave me $500. What do I do? What does he want? And how do I tell him 'no,' because now he's paying my rent.'... Unlike a traditional job... I have thousands of [bosses]." – C17

Easy access & engagement from a variety of attackers. Creators in our study explained that the public-facing nature of creating *and* the ease with which audiences can engage with them exposed them to a variety of attackers. This applied to those with relatively small to quite large audiences:

> "If you put yourself out there, once you reach enough people you're going to have [attacks] happen... I think it's inevitable no matter what it is that you're doing." – C12

Many participants recounted experiences with *parasocial audience members* who felt they "knew" the creators, expecting creators to reciprocate their attention. Such attackers would

request that the creator respond to their comments, engage in phone calls, or gift them products the creator had received. Attacks from parasocial audience members sometimes resulted in unwanted, unsafe physical-world interactions. Participants emphasized the one-sided nature of these relationships:

> "If someone's been watching for a decade or so, they have a lot of insight to our family and personalities... They almost feel like they're friends or they know us, or they're even family. What they don't understand is that we don't know who they are, what their motives are, and it's a very one-sided relationship. So they're sometimes disappointed that you don't just accept them with open arms and have this warm relationship that they expect." – C2

Creators in our study described rarer attacks from *coordinated online mobs*—usually from outside of their audience—potentially instigated by another creator or hostile community. Such attacks are referred to as "raiding" or "dogpiling" [26].

> "[Another creator] made a hate video about me... and it just snowballed... They were bullying me so hard that their subscribers were coming to my [content] and bombarding my comment section. They were coming to my [other platform]. I had to block everything... I just had to stop everything because they were making these terrible videos about me." – C21

Massive popularity and virality further amplify risk. While all creators in our study had similar digital-safety experiences due to their prominence, attacks were described as being more prevalent among creators who had massive reach or experienced viral popularity. Participants believed that the exposure beyond "friendly" audiences that came with broad reach or virality led to a higher risk of experiencing an attack:

> "Anything viral, someone's going to have a problem with it. You could be donating to charity. Someone is going to have a problem with it. You could be saving 500 puppies. Someone is going to have a problem with it." – C14

Participants explained that rapid growth required them to make digital-safety decisions quickly and without sufficient resources:

> "I had 50,000 comments alone on [a piece of content]. It really went viral... I think I just got scared because it was my first time and it blew up so fast." – C6

> "You can gain a huge following without the money to go with it to make your life more secure." – C13

4.2 Social norms

Alongside prominence, creators in our study had to constantly navigate the diverse social norms of their broad—and sometimes unanticipated—audiences as well as the platforms they operated on. Perceived deviations from these norms could result in attacks from audiences and formal action by platforms.

Norms violations of diverse audiences were hard to avoid. Participants said that any—even seemingly benign—content could trigger attacks, indicating that wide-ranging audience interpretations were an important part of risk. C7 shared how a video about making salsa triggered "brutal" attacks:

> "We have videos that show how to make a salsa or hot sauce, and there's full-on hate comments. Because someone doesn't think our salsa is authentic enough. Or their grandma told them never to put this ingredient in salsa. It gets brutal." – C7

Participants' accessibility could lead to them having very broad audiences with incongruent norms. This was particularly challenging during livestreaming, as creators had to process and modulate audience interactions in real-time:

> "[With] livestreams, ... I don't always think through my responses... when you're dealing with [an audience] with diverse backgrounds, diverse personalities, diverse political views, it seems like there's always going to be something that can be wrongly interpreted." – C9

Platform norms violations had a steep cost. Apart from the risk of upsetting (and potentially losing) audience members, creators in our study were especially cognizant that they could lose platform access if their content were framed as violating platform norms or requirements—something offended or even malicious audience members did to harm creators—which would ultimately sever them from their audience or monetization.

> "My biggest concern is the idea that I can slip up or say the wrong thing or something can be interpreted incorrectly, and I can just be gone or banned or censored." – C9

All of these risks influenced the content that creators in our study shared, affecting their voice and presence online.

4.3 Marginalization

Aspects of participants' identities—such as their race, ethnicity, gender, sexuality, age, religion, or physical characteristics—often intersected with their prominence, augmenting the attacks they experienced. While creators in our study universally shared examples of toxic comments about arbitrary aspects

of their on-screen appearances, several shared attacks that focused on marginalized identity or physical characteristics, such as (in their words) "Black," "Jewish," "gay," "old," or "female;" or "overweight," "skinny," or "bald."

Attackers focused on identity characteristics. A participant who self-identified as Black recounted attacks they experienced due to their skin color, as did a fellow Black creator:

> "A gamer told me that they used to have to wear gloves when they were gaming because they didn't want anybody to know they were Black. Then once they took their gloves off, it was a big deal. Somebody saw their hand. That hurt me to the core. That hurt me for society." – C-Anon

Another participant—who self-identified as a gay man—shared that while he had not personally experienced identity-based attacks to the degree of other creators, it was only a matter of time:

> "I can go into [the] comments and see people commenting ... homophobic slurs. I don't understand how I've been so unaffected by something that I see so rampant in other places... it's almost as though I spend my time as a creator waiting for that other shoe to drop." – C-Anon

Intersectional identities compounded risk. Another creator in our study described pervasive toxicity based on intersecting marginalized characteristics—her weight, gender, race, and presentation—and the emotional toll it caused:

> "I'm a woman, I'm overweight. I'm for lack of a better word a b****. I'm just assertive honestly... Hate and harassment is definitely really hard, and then obviously being a woman or being overweight, like there's always going to be something that someone says to me... I hurt because of the audacity of the people—you don't know me." – C-Anon

By operating as public figures, creators' identities were often visible to attackers, exposing participants to identity-based attacks at a large scale.

4.4 Access to a sensitive resource

Participants' ad-based income and influence over audiences—both of which are sensitive resources—increased their risk, often from financially-motivated attackers. Creators in our study explained that attackers frequently targeted their platform accounts to redirect their revenue to the attacker, or to attack their audiences with phishing, scams, or malware:

> "Within the last six months, I've gotten probably three separate emails of people pretending to be a [major brand] manager... They'll find a real person who you can search for on [professional site]... They'll always have a [malicious] link or something to download... I've almost been fooled. They've gotten much better than they were even a few years ago." – C15

As a participant's portfolio and community was tied to their platform accounts, these accounts—and by extension, the creator or staff members who had access to the accounts—were at heightened risk. We unpack this further in Section 5.3.

4.5 Why people create

As our findings on risk factors show, participants were aware of and had personal experience with many digital-safety risks associated with creating. Yet they all felt that the benefits of creating outweighed the risks[3].

> "For the one person who says, 'I had the worst year of my life, but your show got me through it,' that makes the hundred 'You're fat and I hate you' [comments] mean nothing. If you're making a difference in a person's life ... it's all worth it." – C5

All of our participants described benefits that motivated them to create despite the risks, the most common being to express their personal interests, build a community, and monetize their online presence.

> "I started [creating] because there are a lot of topics that I don't have anyone in real life to talk to about. [With creating], I have a place to express my feelings." – C11

All creators in our study relied on the monetization of their content (e.g., ads, subscriptions, sponsorships), though creating was not the primary source of income for all of them (e.g., some had another full-time job). Notably, many participants emphasized that building a community around their content—via platform affordances such as livestreams, comments, chats, and reactions—was an important and valuable part of creating.

> "For me, being a creator means having a more intimate connection with your audience than being on TV... Right after you post content, people can comment and you can talk back to them. Comments are so valuable, so important to building relationships." – C6

These results suggest that creating offers benefits that motivate people to persevere despite the risks. Thus technologists can help by focusing on understanding the risks creators face and improving safety for creators.

[3] Note that all participants were creators at the time of our study; we did not interview people who had been, but were no longer, creating.

5 Digital-Safety Experiences & Concerns

The intersectional risk factors that creators in our study faced collectively impacted the digital-safety harms they experienced or were concerned about. *Prominence* online, *social norms* of broad internet audiences, *marginalization* of their identities and physical characteristics, and access to their community as a *sensitive resource*, combined with platform affordances, to influence participants' perception of safety threats. As such, we present safety concerns as a collective result of risk factors and platform affordances, rather than connecting specific risks and harms. Participants' broad set of threat models was informed by their own experiences and those of other creators. Concerns spanned issues relating to emotional safety (e.g., bullying, trolling), physical safety (e.g., doxxing, stalking, physical violence), relational and community safety (e.g., impersonation, rumours), and financial safety (e.g., scams, account takeover), aligning with the harms framework from Scheuerman et al. [20]. We provide salient examples of attacks across each of these themes to provide a nuanced account of participants' safety concerns.

5.1 Emotional safety

Attacks involving a creator's emotional safety were the most prevalent for participants, often in the form of toxic comments including bullying, sexual harassment, and trolling. Our prior survey estimates that 70% of creators deal with attacks like these sometimes, often, or always [27]. Creators in this study rarely anticipated the emotional toll attacks like these incurred:

> "Just people out of nowhere... criticizing how you look, how you speak, what kind of person you are—in very, sometimes brutal, ways... One comment can just destroy someone's self-esteem." – C10

Compared to other safety concerns, participants thought emotional safety was something they could largely manage themselves—with effort and time—by developing a thick skin (discussed more in Section 6). However, the cumulative impact of attacks and collateral harm still felt challenging to manage.

Cumulative impact of sustained attacks magnified harm. Participants expressed how they had to contend with the collective weight of smaller attacks over time, which could be more harmful than a single severe attack. This was exacerbated, in part, by creators in our study having to maintain an engaged presence with audience members: their role offered no alternative other than triaging or moderating toxic comments, and coping with the resulting emotional harm.

> "The cumulative effect of whatever is going on in any given day—if you read enough comments—is just brutal... In the moment, I'm fine. I try to... make sure I filter that stuff out. But it sticks with you." – C7

One creator who dealt with an extended period of toxic comments and negativity expressed their lack of meaningful alternatives to dealing with sustained attacks:

> "Have you ever worked in a job you don't like where you just don't want to go to work anymore and you wake up and you go 'I don't want to do it'? It's that... But you can't really get away from it because what? Delete your [account]? That's your choice? Just disappear and stop creating entirely?" – C17

Attacks on family & collaborators were especially difficult. Many participants emphasized that they had become inoculated to comments that targeted them personally. However, toxic comments about their family, collaborators, or other relations caused emotional distress, and showed how attacks could impact others who may not be prepared.

> "I posted a video where my [child] walked by in the background. Someone who... had been positive about my videos, said 'you need to put your [child] on a diet.' That just enraged me." – C13

5.2 Physical safety

While most participants had not personally experienced an unsafe physical-world interaction, physical safety was a top concern for all participants. Scenarios they considered involved an attacker discovering their location—perhaps through doxxing or by unintentionally revealing something in their content—resulting in surveillance, stalking, or physical-world harm to the creator, members of their household, or their home.

> "My biggest [concern] is being doxxed and having someone show up at my house." – C6

A small number of creators in our study had been stalked, surveilled, or had audience members show up—uninvited—at their home:

> "I had a major incident that really changed everything for me... I started dealing with a stalker situation. I received a letter to my house—handwritten—from somebody who obsessively watches all my videos. They found out where I lived and everything about my family—where we go, schools, everything... It was so scary for me." – C20

> "You don't know... if they're trying to harm you or... why they're at your house. So it's very uncomfortable. I want to be polite and nice and respectful, but my primary concern is protecting my family." – C2

Privacy losses were irreversible and could lead to physical-world harms. Participants explained that if their personal

information—e.g., residential address, real name—became public, they perceived it as irreversible, exposing them to potential harm in the physical-world. This was common for those who started out with minimal privacy concerns due to their small audience, but who had grown in popularity. For others, information was public because of their other professional careers. C1 shared an incident where their personal information was collated and broadly redistributed via doxxing:

> "[An attacker] started posting my name, my child's name, my [spouse's] name, all our addresses, our phone number, and stuff like that on [platform 1]. And they were trying to post it on [platform 2], comments on [platform 3], everywhere on our website and forums." – C1

C1 tried to make it more difficult for the "average person" to find their information, but felt it couldn't be completely fixed:

> "Nothing's perfect. There's so much public record [data] out there, you're not going to take care of it all. But if you make it a little harder, maybe the next [attacker] won't be able to find you so easily." – C1

Location can be exposed in surprising ways. Creators in our study considered their location to be highly sensitive information that, if leaked, created physical-safety risk. They shared a litany of ways their location might be inadvertently leaked through their content: residential addresses on packages, street signs in backgrounds, reverse image searches of their home's interior or exterior, or seemingly innocuous information being chained together over time to glean more about a creator than was intended.

> "[A creator] told me how [their] house had appeared in a video. One of [their] viewers was a former [law enforcement] agent who took a screenshot, uploaded the image to Google Images, and [the creator's] house showed up because it had been for sale on the MLS[4], which included the address." – C8

Creators in our study discussed vigilance about the risk of content leakage—such as scanning comments for potential personal information—and the mental tax this incurred.

> "[It's] constantly thinking that people are talking about you. Doxxing you. Sharing your personal information that you didn't share publicly... It's tiring and taxing." – C18

This vigilance was learned over time. Participants expressed concern or regret about personal information they had shared prior to an attack, often when their audience was smaller. Such information had usually been shared due to a desire to be authentic and connect with their audience—but in most cases, that information couldn't be taken back later.

Friends and family can be responsible for leaks. Some participants were concerned that their identity and activities might be leaked by family, friends, or peers who may not understand the risks associated with the creator's online presence.

> "A lot of people know me from my previous career: former employees, former co-workers. That scares the living hell out of me. Those people know my true identity... They see me and they're like, 'Whoa, that's [the creator].'" – C18

> "Real life can start spilling details that I've been conscious to cover up. Suddenly, it can be 'I work with [creator] at this [company] on this street in this town. They drive this color of this brand of car.'... I'm equally as careful in real life as I am online to make sure that [physical-world connections and online connections] don't intersect. " – C16

5.3 Relational & community safety

While less top-of-mind compared to emotional or physical safety, creators in our study expressed concern with protecting their relationships and reputation, and by extension, their reputation for maintaining safe communities.

> "I want my audience to feel that sense of community. I want to be able to respond to them. I want it to be a safe space because we talk about very sensitive topics" – C9

Participants emphasized that relational damage was hard to reverse. Attacks that damaged a creator's reputation or their community's safety often involved toxic comments, impersonation, rumors, and/or conspiracy theories. While participants might be able to mitigate further damage once they become aware of the attack(s) by moderating comments or blocking fake accounts, as C4 put it, the "damage is already done."

Attackers target creators to scam their audiences. Creators in our study recalled incidents where scammers impersonated them in order to defraud their audience, for example pushing malware or cryptocurrency scams. C20 shared an experience where an attacker created a homoglyph of their name and reused their profile image to masquerade as the creator:

> "[My impersonator will] reply to [audience] comments with a cryptocurrency ad and a phone number... [My audience] thinks it's me. Then [my audience would say], 'I called the number and you weren't there'... It's like a whack-a-mole." – C20

[4] MLS (Multiple Listing Service) is used by real estate brokers to list properties for sale.

These attacks were not always financially-motivated. They included attempts by attackers to subvert the creator's elevated role within a community to incite conflict. For example, C1 recalled an incident where attackers created fake accounts impersonating C1 to post incendiary comments and chats in a livestream—and later across multiple platforms—to destabilize the creator's community and harm the creator's reputation.

> "They started going around and creating [platform] accounts with my full name on it... going on [another platform] and posting different things there and going around the different communities that we're a part of, with our real names, trying to just cause problems for us... It's nothing major, but it's annoying." – C1

Community safety requires active management. Beyond scams, toxic comments targeted at participants' audiences were common and required constant maintenance:

> "When you get a channel to a certain size, the comments section can be a dumpster fire... I don't want people coming in and harassing another viewer or audience member because they didn't agree with something they said. My main issue has always been: 'How do I make sure I'm taking care of my house?"' – C7

These examples highlight the vigilance needed to keep communities safe (further discussed in Section 6) and the potential for irreversible harm if left unattended even briefly.

5.4 Financial safety

The final safety theme from our study is financial, that is, the risk that creators will lose access to their ad-based or other revenue streams. Creators in our study emphasized that the attacks—such as account takeover attempts by hijackers or false reporting to platforms that resulted in temporary suspensions—left them with few avenues for recourse or recovery. Financial-safety risks were exacerbated by the fact that creators in our study relied on multiple, interconnected platform accounts to drive growth, thus creating a dependency chain that could be disrupted with the loss of a single account.

Platform access is a single source of failure. Creators in our study were particularly cognizant of how account takeovers—and to a lesser extent, denial of services attacks—could sever their platform access. C22 recounted an incident about a fellow creator whose account was hijacked and then held for ransom. C22 reflected on how "devastating" it would be to their livelihood:

> "[A creator] got hacked and couldn't get into [their account]. [They] had millions of subscribers. That would be devastating. You work so hard to build this and then you can't even get in [to your account]... It was a few days [they lost access] but, they did [get the account back]... [The attacker] wanted money to give it back." – C22

C3 experienced a denial of service attack that kept them offline for several hours:

> "[Somebody I met online] tricked me into leaking my IP address... once he got it, he DDOSed me... My whole home internet was shut down for like six to eight hours.... If they know that you're a [creator], and they have your IP address, they're definitely gonna hit you offline." – C3

These attacks highlight the necessity of strong account security practices and robust recovery options.

Abuse-reporting tools can be weaponized. Participants were concerned that abuse-reporting tools—which allow anyone to report potential abuse to a platform—might be misused to remove them from a platform. For example, C21 described multiple instances where other creators tried to get C21 banned from a platform; these other creators asked their audiences to report C21 for abuse, posting instructions on how to do so:

> "[They] showed people how to report my [account]. It was false because I wasn't doing anything. They reported me for bullying.... [They'd say,] 'Let's get [creator's account] taken away."' – C21

Absent a point of contact at the platform—e.g., an account manager available to large creators—participants did not know how to reverse platform decisions they found to be erroneous.

6 Protective Practices

Creators in our study employed protective practices such as limiting what they shared, moderating comments, or seeking external support in response to their safety concerns and experiences. We discuss the most salient protective practices here; additional practices are covered in the Appendix.

6.1 Adopting protective practices

A majority of creators in our study began their journeys without a realistic expectation of the risks associated with prominence, social norms, or their access to sensitive resources—and thus, without sufficient protective practices. A minority started creating with an extensive set of protective practices motivated by a relatively sophisticated understanding of the threats they would likely face, or due to past experiences with marginalization.

A majority started creating with minimal protections. Two-thirds of the creators in our study adopted protective practices *in response to* attacks or concerning situations. In

this reactive framing, it was common for participants to describe a particular experience that caught them by surprise and "crossed a line," prompting them to take action. Attacks that threatened the creator's physical, relational, or financial safety commonly served as catalysts. For example, after C20 resolved a severe stalking incident, they and their partner completely changed who they talked to and what they shared.

> "Now we're both very private. We don't tell anybody anything, because we don't know if there is a mole in our life... I was friends with a few other [creators]. I don't tell them anything anymore." – C20

A minority started creating with strong protections. One-third of creators in our study described adopting protective practices from the onset of their journeys, refining their approach in conjunction with their rise in popularity. Common rationales for early protections included wanting to keep their career (and identity) secret from the start; receiving early advice from other experts (discussed shortly); as well as an early awareness of the risks of being a public figure on the internet. For example, one participant (who self-identified as Black) reflected on their wide-ranging mental model of threats, influenced in part by their past experience with marginalization:

> "I've been on this planet for a long time. It's sort of just common sense." – C-Anon

For some, these practices were motivated by negative experiences prior to creating, such as C14, who had experienced sexual harassment and was primed to consider risks related to stalking and surveillance. Nevertheless, they believed they had to keep improving their protections to avoid possible attacks:

> "When it's a much bigger scale on the internet... I just wanted to double down on this before it could ever even be a problem." – C14

Access to trusted experts or advice was crucial. Having early access to security experts or advice played a crucial role in whether creators in our study established robust protective practices. For example, one participant—who had a background in IT—had a practice of rotating their passwords every 90 days, using multi-factor authentication, custom encryption, and more. Another participant had friends and family in law enforcement who helped them prepare. A third participant had access to a playbook of security advice as a result of other professional activities.

Another important source of expertise came from other creators; about half of our participants learned about protective practices from creators they knew, either indirectly—e.g., observing how other creators responded to attacks—or via close connections with creators who directly shared advice.

Unfortunately, most creators in our study did not have access to sufficient expertise and (as noted above) were surprised by and felt under-protected for attacks early in their journeys. We further discuss the imperative of sharing such expertise with creators who lack similar access in Section 7.

	Distancing	Moderation	Account security	Network security	Physical security	External support
Emotional safety	●	●	×	×	×	●
Physical safety	●	●	●	×	●	●
Relational safety	×	●	×	×	×	●
Financial safety	×	×	●	●	×	●

Table 1: Practices adopted by creators in our study and the types of safety they supported (marked by dots). Details around account security, network security, and physical security—which rarely differed from other at-risk populations—are in the Appendix.

Creators developed a "thick skin". Creators in our study discussed having to reset their safety expectations, accepting the elevated risks they faced. They developed a "thick skin" so that attacks would not become emotionally exhausting. This thick skin was deliberately built and maintained over time and with effort. Some participants were determined to not let attackers "win" and impact their creating, despite feeling harmed by attacks. However, creators in our study were clear that some harm was unavoidable:

> "I realized that... I cannot take it personally because... it really doesn't have anything to do with me... So it's that realization that helps me move forward, even though it hurts sometimes. Even to this day, if you get a comment... it can throw you into this cycle of... negative thoughts." – C10

6.2 Distancing to maintain privacy

Creators in our study responded to threats involving their emotional and physical safety by adopting distancing behaviors to improve their privacy—such as self-censorship or avoiding platform features—which was at odds with audience expectations of their authenticity and accessibility, and their ability to build prominence (see Table 1). However, participants expressed that it was difficult to predict when distancing would be helpful (e.g., what information to avoid sharing), and impossible to "undo" sharing on the internet.

It's difficult to predict what shouldn't be shared. A majority of creators in our study reported self-censoring—limiting

what they shared about themselves—as a core protective practice, often to protect against attacks like content leakage and toxic comments. Examples included participants who used pseudonyms across their online presence to protect their real name; avoided controversial topics; purposefully shared fake hints about where they lived; used P.O. boxes (sometimes in cities where they didn't live); avoided sharing their current location; never uploaded content until they departed from where they filmed; or kept the backgrounds in their photos and videos free of any identifiable features.

However, the unpredictable nature of what content might trigger or augment an attack meant that participants often adopted self-censorship after experiencing an attack or learning about a new risk. For example, C13 was more cautious about sharing their location after receiving threats of violence:

> "[A commenter] was actually making threats to my personal safety. That's where it fully crosses a line... I don't ever tag location in any content that I'm doing or talking about it... Like I'm at dinner at a certain place, I'm just very aware of not posting that until I've already left so that I don't have to worry about things happening." – C13

It's difficult (and often impossible) to take something back. Half of our participants reported retroactively taking action to remove personal information from the internet after experiencing or becoming concerned about attacks. Some deleted or edited old content that revealed personal information. Others turned to internet directory scrubbers to help with the proliferation of personal data at scale. However, they found it difficult to fully remove information from the internet once it had been posted, thus hyper-vigilance was required to prevent mistakes:

> "[You need to] triple check your videos to make sure there's nothing personal—like phone numbers or your name or anything else—because once stuff gets out, its very difficult to delete." – C15

Severe attacks led to reducing footprints & leaving platforms. At the extreme, some participants reported abandoning features—such as direct messages or livestreaming—and even abandoning platforms, at least temporarily if not permanently, in response to physical-safety threats or serious emotional harm. For example, C21 described severe toxic content attacks they experienced across multiple platforms that led to their abandoning a platform and their content:

> "I left the platform for five months. I thought I would never come back. I was so scared. Then I deleted all of that content. I pretended I had never posted anything. I was trying to build my name up again, because when you searched for me, all that would come up was terrible things." – C21

The same creator shared that once they returned, they chose to avoid livestreams—even though this decision had financial repercussions—because they could not control the risks:

> "I made a pretty good amount of money [from livestreams], but I had to just sacrifice that because it's not worth my [well-being] to do it... I'm too afraid. I can't control who comes to my livestreams." – C21

These findings on well-being and financial trade-offs add depth to our previous research that found 44% of creators left a platform temporarily due to a variety of hate and harassment attacks, with 19% leaving permanently to avoid attacks [27].

6.3 Moderation for safety

To protect themselves and their audiences from a wide range of safety concerns, nearly every creator in our study reported engaging in some form of comment or chat moderation (see Table 1). Participants felt empowered by having these broadly applicable safety tools, though some cited confusion about how some options worked, acknowledged a lack of awareness about some options and tools available to them, or recognized the limited reach of their control over some types of attacks.

Keyword filtering, reporting, & blocking were crucial *within* a platform. Creators in our study set up keywords to automatically filter toxic content or personal information such as their real name (if the creator was anonymous or pseudonymous), the city they lived in, or other identifying information about themselves or their families they deemed to be sensitive. To mitigate concerns about scammers targeting their audience, participants used these systems to moderate any links shared via their community. Automated filtering was also used to mitigate attacks by bots and raids. Other techniques included manual reporting, hiding, or blocking abusive content or community members that they deemed could cause harm.

> "We try to make sure that we have a high level of filtration on... [to prevent] either spam or harassment to... another viewer or targeted audience member." – C7

Differences in how moderation features worked and the terminology used by different platforms confused some creators in our study. They wanted to better understand the scope of moderation actions such as "hide," "delete," and "block," including who would be impacted by, and who could see, the outcomes of these actions. Another frequently cited limitation of these tools was the ease with which attackers could migrate to new accounts—or other platforms outside the creator's sphere of influence—thus circumventing participants' moderation actions. Additionally, attackers could flaunt moderation actions as a way of reinforcing their reputation:

> "If you block someone on [platform], it flat out tells the person, 'So and so blocked you.' And then people

use it as a badge of honor. It's weird. They'll take a screenshot and show everyone." – C13

Despite these limitations, participants reported that existing moderation tools were critical in mitigating immediate forms of harm and maintaining their community's culture and safety.

Norms & moderators help, but also create risk. One way creators in our study helped keep their communities safe and protect themselves from relational damage was to establish community norms and instate moderators to help uphold these norms. This was most commonly reported by participants who livestreamed or had discussion-based communities, though some had moderators help with comment sections as well. Nearly all moderators were unpaid, though some participants enlisted paid channel managers who performed multiple responsibilities, including moderation.

Participants sometimes used platform-provided tools to set their community's tone and expectations, such as creator-written community guidelines; restricting comment access to followers or vetted audience members; and permitting their moderators to review, remove, ban, or time-out audience members in comments or chat. Some creators in our study established expectations with their moderators on the types of content that were and were not allowed in their communities.

Despite the wide-spread use of moderators, participants had lingering concerns with the lack of fine-grained access control over moderation features or logs of sensitive actions:

> "I trust my moderators... I have five or six... but let's say one goes rogue... [They] can start deleting a bunch of the top comments on my videos. That would be horrible for my video's engagement... And there's no log of it." – C3

Likewise, moderators became an extension of the creator. Any mistakes made by the moderator reflected back on the creator:

> "I had to direct message [my moderator] and say ... 'Unfortunately, you, as a moderator, automatically now are associated with me. Whatever you say is basically coming out of my mouth'... I can't have someone interpreting what [the moderator] said in a negative fashion." – C9

6.4 External support for serious issues

Creators in our study often turned to formal or out-of-band platform communication or other institutional help when responding to safety issues (see Table 1), especially the serious ones. The core challenge with this category of protective practices was having a point of contact who was willing and able to help. Many participants did not have such contacts.

Platform support helped when creators knew who to contact. Creators in our study who experienced false reporting or account hijacking—which caused relational and financial harm—told us that they reached out to platforms for help. Examples included contacting an ad hoc representative, such as a friend who worked on the platform, or engaging in appeal or recovery workflows to resolve issues—though participants reported that formal recovery options were slow and opaque.

> "[The attacker] reported [my] video to [the platform]... and I got a [violation]... I went back and forth with [the platform] for months about this... [I] got somebody at [the platform] to watch the video, and the [violation] was removed." – C11

However, participants did not think those just starting out would have the necessary relationships to ask for support:

> "If you're a small creator... and someone is trying to doxx you or attack you online and you're not part of a partnership program or you don't have access to an account manager, you might as well whistle in the wind." – C2

Law enforcement & legal aid were only helpful sometimes. Participants had mixed outcomes when they sought help from law enforcement, something about a third of our participants did. Some of these participants successfully coordinated with local police to prevent swatting, a physical-safety concern:

> "I called my local police station, and I told them that I'm a [creator] in the city and I wanted to have swatting protection... [They] set me up so that before somebody calls a SWAT [team] to my house... the police would have to call me first... and verify that something bad is actually happening." – C3

However, not all law enforcement was understanding of participants' needs. One participant explained that law enforcement refused to help with a physical threat from a person they knew the identity of "unless [the creator was] physically threatened in person," leaving the creator feeling "helpless." C20 also expressed friction with law enforcement:

> "[Creators are] still not taken that seriously. It's not like you're an A-list actor or something... You have a social media platform." – C20

Advice from other creators was valuable for those with connections. Creators in our study described sometimes feeling isolated, especially when they experienced attacks. Multiple participants expressed that having creator friends or being part of creator communities helped them find implicit validation and, at times, tips on how to navigate attacks:

> "If you're an army of one, it's a tough army to be in. So when you have a lot of people around you, that sometimes you don't need answers, you just need the ability to vent or hear from someone else that they're going through the same struggles." – C7

For some, these friendships were established in the natural course of creating content—finding others that create similar content, working together, or meeting at events. However, multiple participants had not successfully found a community of peer creators and expressed a need for help doing so.

> "[If] there were a forum for creators where they know it would be a safe place to talk about the issues they're having, and how they handle it. If there were something like that, I would have known a lot earlier that you can block people without them knowing it." – C13

7 Discussion

As our results show, the risk factors that creators have shape their digital-safety experiences across a multitude of platforms. We have shown how these factors combine to create a different risk profile than other populations, expanding our understanding of creators and at-risk users more generally. Additionally, we explore recommendations that could help creators, which include directions for increased agency, platform support, and safety advice for creators at the start of their journeys.

7.1 Creators' unique safety considerations

All creators in our study existed at the intersection of the risk factors *prominence*, *social norms*, and *access to a sensitive resource*; many also experienced *marginalization* (depending on their identity or physical characteristics), which amplified their risk of digital-safety threats. Of recently studied at-risk user populations in Warford et al.'s review [29], none were known to have this combination of risk factors—making creators a novel population to study for their unique intersection of digital-safety risks and needs.

Four populations in Warford et al.'s review shared two risk factors (*prominence* and *access to a sensitive resource*): activists [1, 7, 25], people involved with political campaigns [6], journalists [18, 19], and NGO staff [5, 15]. However, the *social norm* for creators to provide consistent, authentic access to themselves and engage regularly with online audiences amplified the existing risks associated with prominence, access to sensitive resources, and marginalization. Creators in our study also tended to moderate their communities themselves or with small groups of volunteers, often acting as the front line against digital-safety threats. Journalists and these other populations were usually supported by staff or organizations who helped them stay safer online or establish proactive safety strategies. Creators did not yet have similar support systems, and instead relied on platforms and ad-hoc connections between each other, leaving many of our participants without support at times.

Possibly most similar to creators are journalists, who also create content and are often subject to toxicity online [16]. But prior work emphasizes that journalists' top concern was protecting sensitive resources (e.g., sources, stories, and data), prompting use of secure accounts and private, encrypted communications [18, 19]. Similarly, creators in our study expressed a sense of responsibility for protecting both themselves and their sensitive resource—audience members—which incurred an additional operational and emotional cost. Unlike journalists who have other avenues for audience engagement and report infrequently experiencing serious harassment [16], creators' consistent presence and interactions with their audience on social media made them the target of toxic content, information leakage, and targeted attacks—like scams—to exploit their audience's trust. However, changing dynamics in journalism, such as an increased focus on online engagement and harassment concerns, may increase alignment between the two groups.

7.2 Multiple tensions & challenges

While our prior work enumerated how often creators experience forms of hate and harassment [27], this study provides a nuanced depiction of creators' lived experiences and the trade-offs they routinely weighed regarding their digital safety. Using these results, we synthesize tensions in their safety environments through the lenses of harms, protective practices, and risk factors, in order to understand difficulties creators face staying safe.

Nuanced understanding of creators' fears. Applying a harm-based lens illuminated creators' fears. For example, while we previously found that experiences with physical-safety threats were uncommon [27], concerns about physical safety were strongly felt by the creators in our study, and often discussed as a top safety concern. They felt there was little they could do to protect against potential information leakage and resulting physical safety concerns: once information was on the internet, it was difficult—or even impossible—to remove, and even unintentional sharing of minor details (as a result of authentic content creation) could increase their risk in unpredictable ways. Downstream impacts included relational harms that felt difficult to reverse and demanded constant attention, and potential financial harms as a result of coordinated attacks or inadvertent violations of platforms' guidelines.

Limits of current protective practices. Our results on creators' protective practices revealed tensions between what they could control and where they needed more support to meet existing safety needs. Early in their creator journeys, participants wanted to be authentic online, despite most not having the expertise to do so safely. These early sharing decisions created privacy and physical safety concerns as participants' audiences and risk awareness grew. But decisions to disclose information were hard to reverse, leaving creators in our study feeling unable to prevent some attacks, due to a lack of knowledge or ability to mitigate them. Some participants sought external support from platforms and law enforcement to bridge

this gap. However, some approaches, such as reporting, felt opaque and ineffective—something highlighted in our prior work [27]. These findings suggest that creators might benefit from research into improving safety and support systems, including reporting on platforms and beyond.

Trade-offs with growth and safety. The risk factors that shape a creator's risk profile can inherently lead to tensions. Growing their prominence and increasing their financial success (often important goals), increased the frequency of attacks and number of attackers. This was especially true of viral content: while it gave creators access to critical new audiences, it simultaneously included unpredictable social norms and hostile audiences irrespective of the type of content a creator uploaded. Further, audiences expected creators to constantly engage (e.g., in comments, live streams, chats, and more) even when a creator might otherwise step away for self care. Taken as a whole, creators had to navigate when to enact protective practices, such as privacy distancing behaviors, and how those practices would interfere with their growth and potential financial opportunities. Researchers proposing new protections or coping practices should take these tensions into consideration.

7.3 Towards solutions

While some creators in our study felt these risks to their digital safety were inevitable—and that there was little they could do to better protect themselves—most had ideas, suggestions, and hope that solutions could alleviate their risks.

Digital-safety resources. Creators in our study thought they would have benefited from curated digital-safety resources at the start of their creator journeys, as well as after they experienced an attack. Per our analysis, a majority of participants did not adopt protective practices until after experiencing an attack that "crossed a line." They described retrospectively feeling ill-informed about the types of attacks they would eventually experience after becoming more prominent. Our results also showed that creators in our study could not always predict the irreversability of decisions to share personal information (like their name or home location) or the general kinds of protections they should employ. These results suggest the importance of outreach to new creators that informs them of potential digital-safety risks associated with creating and how to protect themselves from the onset, as well as advice for how to react to incidents.

Creators in our study were concerned that family and friends who lacked an understanding of the risks creators face could inadvertently leak personal information or be targeted as stepping stones toward harming the creator. Some participants discussed having to extend their protective practices to their immediate relations, such as paying to scrub personal information from the internet for those they lived with, or explicitly telling family and friends to not speak to people about their work as a creator. Educational resources for friends and family of creators might take the form of basic privacy and security training, focusing on raising awareness of the risks they and creators may face and instructing them on relevant protective practices. Security educators could help by building and testing what resources are most helpful for early stage creators, their family members, acquaintances, and active audience members.

Creator community building. Many participants said they would benefit from ways to develop communities of peer creators. Participants expressed that peer connections were valuable tools for emotional support, gathering digital-safety advice, and sharing tips on how to grow and succeed. But forming these connections was ad-hoc for participants, and multiple wanted help connecting with peers. However, creators in our study noted that these relationships are best when developed organically and they had a general concern that platform-organized community building could feel contrived and inhibit the authentic connections they found valuable. Developing ways for creators to organically support each other and share practical safety advice could help foster individual and community resilience in the face of attacks. These community building efforts should consider and mitigate adversarial or competitive motivations, as our results also showed that creators sometimes attacked other creators.

Expanding platform-provided tools. Creators in our study frequently relied on platform-provided tools, such as moderation, to help protect themselves and their communities from attacks. Participants noted areas in which moderation solutions could be improved. They were confused by the language describing some features, such as the difference between "blocking" and "hiding," especially as these terms were not uniform across platforms. Usability studies could explore how to clarify these features for users. Creators in our study also expressed an interest in more granular moderation controls, such as the ability to customize which controls they could delegate to moderators (e.g., enabling a live chat moderator to only time out audience members, but not ban them). Some participants also wanted community moderators to have increased access to privacy and security best practices, since a moderator's access to a creator's community members also put the moderator at higher risk of attacks.

Creators spent considerable effort moderating their community and managing their digital safety, and would benefit from any platform support that could make this easier. Creators in our study recognized that moderation tools were limited to protecting their immediate communities on each platform; they did not have control over other communities within a platform, let alone across platforms. As threats and platforms evolve, the security research community should continue to invest in improved moderation tools that can provide creators reasonable agency beyond their communities and across multiple platforms. However, creators in our study also expressed the emotional and operational toll of moderating for themselves

across their growing online presence, suggesting a need for design solutions which scale and reduce these burdens. Beyond moderation, the research community could invest in technologies or other protective measures for attacks that creators felt more helpless towards due to scale or novelty, including the removal of personal information that has already been shared and newer attacks like false reporting.

8 Research in Action at YouTube

In response to our research—alongside ongoing user experience research and insights from the Creator in Residence program [33]—YouTube has created a growing set of safety-focused education and features. For example, YouTube launched its Creator Safety Center in 2022 [32], which provides education on protective practices creators can use to prevent, cope with, and recover from digital-safety issues. YouTube also supports creator-to-creator knowledge sharing on digital safety via series like YouTube Reframe [35].

Beyond educational resources, YouTube offers moderation tools—which have expanded since this research was conducted—including blocked word lists and the ability to block or hide abusive users, hold comments for review, and delete comments. For live chats, creators can put users in time out or limit how often they can post. To scale comment moderation, creators can increase the strictness threshold for the automatic detection of inappropriate comments that are held for review. YouTube creators have access to two-factor authentication (required by default), Advanced Protection [11], and a host of account security protections [10]. If incidents occur, creators can report content or other users to YouTube. Creators who follow certain eligibility steps to join the YouTube Partner Program [34] also have direct access to Creator Support teams for help. Finally, YouTube brings together cohorts of creators to improve community building as part of its Creator in Residence program [33]. In sharing these lessons, we hope to empower other online creation platforms to help ensure the digital safety of participants.

9 Conclusion

We interviewed creators about their experiences with digital-safety threats, protective practices, and recommendations for how to overcome the digital-safety challenges they experienced. We found that creators in our study had encountered and were concerned with a wide array of digital-safety threats across all the platforms they participated on. Risks—including to their emotional, physical, relational, and financial safety—stemmed from creators' prominence online, the diverse norms of their broad audiences, and their ability to steer the attention of a large audience through consistent engagement. For some, these risks were compounded by marginalized identities or characteristics. Creators employed a variety of protective practices—including distancing, moderation, and seeking external or social support—though they tended to adopt these practices only in response to an attack or a concerning situation. Only some of the creators in our study were aware of the risks they might potentially face when beginning their journeys as creators. However, most of the creators reported wishing they knew about the risks and how to protect themselves earlier. As barriers to creation lower, we believe that improving digital-safety technologies and education for the heightened risks that creators face can help to elevate the digital-safety of *anyone* who posts content online.

Acknowledgements

We thank the anonymous creators who participated in our study and shared their personal stories. This research would not have been possible without your engagement and desire to make the ecosystem safer for all creators. We also thank our anonymous reviewers for their feedback and suggestions which helped to improve our work.

References

[1] A. Alvarado Garcia, A. L. Young, and L. Dombrowski. On making data actionable: How activists use imperfect data to foster social change for human rights violations in Mexico. *PACM HCI*, 1(CSCW):1–19, 2017.

[2] C. Barwulor, A. McDonald, E. Hargittai, and E. M. Redmiles. "disadvantaged in the american-dominated internet": Sex, work, and technology. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.

[3] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101, 2006.

[4] K. Chayka. What the "Creator Economy" promises—and what it actually does. https://www.newyorker.com/culture/infinite-scroll/what-the-creator-economy-promises-and-what-it-actually-does, 2021.

[5] C. Chen, N. Dell, and F. Roesner. Computer security and privacy in the interactions between victim service providers and human trafficking survivors. In *Proceedings of the USENIX Security Symposium*, 2019.

[6] S. Consolvo, P. G. Kelley, T. Matthews, K. Thomas, L. Dunn, and E. Bursztein. "why wouldn't someone think of democracy as a target?": Security practices & challenges of people involved with u.s. political campaigns. In *Proceedings of the USENIX Security Symposium*, 2021.

[7] A. Daffalla, L. Simko, T. Kohno, and A. G. Bardas. Defensive technology use by political activists during the sudanese revolution. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2021.

[8] S. Eckert. Fighting for recognition: Online abuse of women bloggers in germany, switzerland, the united kingdom, and the united states. *New Media & Society*, 20(4):1282–1302, 2018.

[9] B. Finley. Popular YouTuber Dream has been doxed. https://gamerant.com/dream-face-reveal-youtube-dox/, 2021.

[10] Google. Account Security. https://myaccount.google.com/security, 2023.

[11] Google. Advanced Protection Program. https://landing.google.com/advancedprotection/, 2023.

[12] V. Hamilton, H. Barakat, and E. M. Redmiles. Risk, resilience and reward: Impacts of shifting to digital sex work. *arXiv preprint arXiv:2203.12728*, 2022.

[13] T. Hatmaker. Twitch sues two users for harassing streamers with hate raids. https://techcrunch.com/2021/09/13/twitch-hate-raids-lawsuits/, 2021.

[14] S. Jhaver, Q. Z. Chen, D. Knauss, and A. X. Zhang. Designing word filter tools for creator-led comment moderation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2022.

[15] S. Le Blond, A. Cuevas, J. R. Troncoso-Pastoriza, P. Jovanovic, B. Ford, and J.-P. Hubaux. On enforcing the digital immunity of a large humanitarian organization. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2018.

[16] S. C. Lewis, R. Zamith, and M. Coddington. Online harassment and its implications for the journalist–audience relationship. *Digital Journalism*, 8:1047 – 1067, 2020.

[17] A. McDonald, C. Barwulor, M. L. Mazurek, F. Schaub, and E. M. Redmiles. "it's stressful having all these phones": Investigating sex workers' safety goals, risks, and practices online. In *Proceedings of the USENIX Security Symposium*, 2021.

[18] S. E. McGregor, P. Charters, T. Holliday, and F. Roesner. Investigating the computer security practices and needs of journalists. In *Proceedings of the USENIX Security Symposium*, 2015.

[19] S. E. McGregor, F. Roesner, and K. Caine. Individual versus organizational computer security and privacy concerns in journalism. In *Proceedings on Privacy Enhancing Technologies*, 2016.

[20] M. K. Scheuerman, J. A. Jiang, C. Fiesler, and J. R. Brubaker. A framework of severity for harmful content online. *Proceedings of the ACM on Human-Computer Interaction*, 2021.

[21] B. Schoon. Youtube will force creators to use 2-step verification on google accounts starting this year. https://9to5google.com/2021/08/24/youtube-2-step-verification-requirement, 2021.

[22] A. Shen. Phishing campaign targets youtube creators with cookie theft malware. https://blog.google/threat-analysis-group/phishing-campaign-targets-youtube-creators-cookie-theft-malware/, 2021.

[23] S. Sobieraj. Bitch, slut, skank, cunt: Patterned resistance to women's visibility in digital publics. *Information, Communication & Society*, 21(11):1700–1714, 2018.

[24] A. Strohmayer, J. Clamen, and M. Laing. Technologies for social justice: Lessons from sex workers on the front lines. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.

[25] B. Tadic, M. Rohde, V. Wulf, and D. Randall. Ict use by prominent activists in republika srpska. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016.

[26] K. Thomas, D. Akhawe, M. Bailey, D. Boneh, E. Bursztein, S. Consolvo, N. Dell, Z. Durumeric, P. G. Kelley, D. Kumar, D. McCoy, S. Meiklejohn, T. Ristenpart, and G. Stringhini. SoK: hate, harassment, and the changing landscape of online abuse. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2021.

[27] K. Thomas, P. G. Kelley, S. Consolvo, P. Samermit, and E. Bursztein. "It's common and a part of being a content creator": Understanding how creators experience and cope with hate and harassment online. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2022.

[28] J. Uttarapong, J. Cai, and D. Y. Wohn. Harassment experiences of women and lgbtq live streamers and how they handled negativity. In *ACM International Conference on Interactive Media Experiences*, 2021.

[29] N. Warford, T. Matthews, K. Yang, O. Akgul, S. Consolvo, P. G. Kelley, N. Malkin, M. L. Mazurek, M. Sleeper, and K. Thomas. Sok: A framework for unifying at-risk user research. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2022.

[30] D. Y. Wohn and G. Freeman. Audience management practices of live streamers on twitch. In *ACM International Conference on Interactive Media Experiences*, pages 106–116, 2020.

[31] YouTube. Creator Research. https://www.youtube.com/creators/research/, 2023.

[32] YouTube. Creator Safety Center. https://www.youtube.com/creators/safety/, 2023.

[33] YouTube. Partners in Innovation: the Creator in Residence Program. https://blog.youtube/inside-youtube/partners-innovation-creator-residence-program/, 2023.

[34] YouTube. YouTube Partner Program overview eligibility. https://support.google.com/youtube/answer/72851, 2023.

[35] YouTube. YouTube Reframe. https://www.youtube.com/playlist?list=PLpmwWuIh57wbu940CXnXf1d9eaJrBQNHw, 2023.

Appendix

Additional Protective Practices

Account security. Creators adopted modest account security measures to address threats to their financial safety (to prevent losing access to their creator accounts via account hijacking) or physical safety (to prevent leakage of personal information via authorized account access) (see Table 1). Most creators relied on two-factor authentication (2FA) to protect their creator accounts—in part due to some platforms requiring 2FA on all monetized accounts [21]. They mostly used weaker forms of either SMS, a code-generating app, or a prompt on their phone. Only one creator explicitly mentioned using a hardware security token (e.g., a Titan Security Key). Some were aware of this stronger option, but reasons for not adopting security keys included not knowing where to acquire one, the cost of the key, fear of losing access, and lack of support by multiple platforms. These secure practices did not always extend to personal accounts, leaving potential gaps in protection.

Roughly half of the creators discussed password best practices, such as using a unique password per website, using a password manager, or using strong passwords as part of their strategy for protecting their accounts.

Network & physical security. For creators who experienced network-based overloading attacks, which harmed them financially by limiting their access to their accounts, anonymity services such as VPNs were common. A subset of creators mentioned other network security measures including firewalls, multiple routers, ISPs that allowed IP changes, and DDoS protection services such as CloudFlare. Regarding physical security—practices put in place for physical safety—multiple creators reported setting up security cameras around their homes. One creator reported purchasing a weapon out of concern for their family's safety, after multiple viewers came to their house; they explained: "We had to arm ourselves. We had to put up gates."