“It’s common and a part of being a content creator”: Understanding How Creators Experience and Cope with Hate and Harassment Online

Kurt Thomas
kurtthomas@google.com
Google
USA

Patrick Gage Kelley
patrickgage@google.com
Google
USA

Sunny Consolvo
sconsolvo@google.com
Google
USA

Patrawat Samermit
patrawat@google.com
Google
USA

Elie Bursztein
elieb@google.com
Google
USA

ABSTRACT

Content creators—social media personalities with large audiences on platforms like Instagram, TikTok, and YouTube—face a heightened risk of online hate and harassment. We surveyed 135 creators to understand their personal experiences with attacks (including toxic comments, impersonation, stalking, and more), the coping practices they employ, and gaps they experience with existing solutions (such as moderation or reporting). We find that while a majority of creators view audience interactions favorably, nearly every creator could recall at least one incident of hate and harassment, and attacks are a regular occurrence for one in three creators. As a result of hate and harassment, creators report self-censoring their content and leaving platforms. Through their personal stories, their attitudes towards platform-provided tools, and their strategies for coping with attacks and harms, we inform the broader design space for how to better protect people online from hate and harassment.

CCS CONCEPTS

• **Security and privacy → Human and societal aspects of security and privacy**.

KEYWORDS

Security and privacy, hate, harassment, creators, content moderation

ACM Reference Format:
Kurt Thomas, Patrick Gage Kelley, Sunny Consolvo, Patrawat Samermit, and Elie Bursztein. 2022. “It’s common and a part of being a content creator”: Understanding How Creators Experience and Cope with Hate and Harassment Online. In *CHI Conference on Human Factors in Computing Systems (CHI ’22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3491102.3501879

1 INTRODUCTION

Online social media platforms like Instagram, TikTok, and YouTube have given rise to a modern creative class—the “creator”. These artists, storytellers, gamers, entrepreneurs, and influencers regularly upload content with the potential to reach millions of viewers. But alongside increasingly lucrative financial opportunities tied to content creation [39, 41, 45], the public spotlight of social media puts creators at heightened risk of hate and harassment. Examples of attacks include Leslie Jones, who temporarily left Twitter after experiencing a wave of racially-motivated toxic comments [59], or Nate Hill, who had emergency authorities called to his residence while live streaming on Twitch [36]. Attacks can originate from audience members—but also competing content creators [4] and coordinated online mobs [43]. This has led some creators to cite burnout as a result of contending with toxicity and harassment [40].

In order to prevent or mitigate attacks, creators and platforms rely on a variety of levers including community guidelines [53], content moderation tools [32, 33, 52, 71], and abuse reporting [14, 34], with an eye towards increasingly automated detection [21, 29]. However, multiple critiques have been leveled at these systems, including fragility to evasion or bias [5, 18, 35] and a lack of transparency for what policies are enforced [8]. Equally challenging, protections for hate and harassment are often disjointed, requiring unique interventions for a creator experiencing toxic comments, versus a creator whose personal information was leaked, or a creator being overloaded with negative reviews and ratings [64]. As such, we argue there is a gap in the community’s current understanding for which solutions best protect creators, and how to prioritize improvements based off the frequency of attacks and their resulting harms.

We inform the design space of protections by conducting the first research study on the extent to which creators experience hate and harassment attacks today and the coping practices they currently employ. We surveyed 135 U.S.-based creators to understand their general concerns with hate and harassment online, their personal experience with attacks (including toxic comments, impersonation, stalking, account hijacking, and more), other parties targeted by attacks (such as their audience, family, or other creators), and any

CHI ’22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04.
https://doi.org/10.1145/3491102.3501879

resulting harms. Additionally, we asked creators what protections and coping practices they rely on, gaps they perceive with existing solutions, and any advice they would offer to new creators for staying safer.

We find that while a majority of content creators view audience interactions favorably, 95% of creators in our study were able to recall an incident of hate and harassment. Roughly 70% of creators reported experiencing bullying, trolling, sexual harassment, and identity attacks *more often than rarely* through the course of their careers. And for 36% of creators, such attacks were a *regular* occurrence. While many creators brushed off attacks and advocated just ignoring haters, 22% of creators intentionally censored what they post—about themselves or their beliefs—to avoid negativity. Another 44% said they have left a platform at least temporarily due to hate and harassment, while 19% reported permanently leaving a platform and the community they had cultivated to avoid further attacks. Paired with modeling and other statistical analysis of what factors correlate with creator's experiencing higher risk of attacks, our findings illustrate the diversity and severity of threats facing creators today.

Many creators turned to platform-provided moderation tools such as keyword blocklists and manual review queues to address hate and harassment, with 82% of creators viewing such solutions favorably. Creators expressed that platform reporting was one of the least effective interventions—with just 55% of creators viewing it favorably—due to a perceived lack of follow through or transparency. Beyond mechanistically triaging abusive content, creators discussed their reliance on friends, family, and other creators to help them cope with the emotional strain of attacks. After asking what advice might help new creators, we also identified a set of proposed best practices—such as having a playbook, how to configure moderation tools, and how to keep personal details private—that creators felt would help mitigate attacks. As creators represent an at-risk population, it is our belief that their lived experiences act as a portent for how hate and harassment will evolve online. Lessons and protective practices that emerge for creators can thus inform the broader solution space for general internet users.

2 RELATED WORK

Our research was influenced by prior work that investigated gaps in community guidelines, content moderation, and reporting—systems that creators often rely on—as well as survey estimates of hate and harassment online. We also situate our work in the broader space of at-risk user research.

2.1 Platform-provided mitigations

Currently, platforms rely on a combination of tools to mitigate hate and harassment including community guidelines, content moderation, and reporting. For example, nearly every major platform includes policy guidelines that explicitly prohibit hate and harassment [53]. Community-based platforms like Reddit and Discord allow operators to extend these rules with publicized community expectations, which also play a role in transparency around moderation decisions [17, 55].

Beyond community governance, built-in moderation tools allow community or channel operators to hide content (though it remains visible to the poster), remove content, or entirely block content. Additional controls include issuing bans, timeouts, or otherwise regulating discussion [70]. For live conversations, such as on Twitch, Discord, or YouTube, platforms provide tools for slowing the rate of messages, or restricting chatting to subscribers or followers [73]. Across all these controls, decisions stem either from manual review; configurable rules based on keywords, regular expressions, or domain reputation for hyperlinked content; or via automated classification systems [31, 70, 73]. For example, YouTube allows creators to manually review and control the visibility of every comment that appears on their videos, customize keyword blocklists, or rely on platform-provided classifiers to flag potentially inappropriate content [72, 73].

Finally, reporting capabilities on sites such as Reddit, Twitter, Facebook, and YouTube include flagging individual comments or accounts in order to trigger review by the platform—in turn improving models of abusive content [14].

2.2 Gaps in protection

A variety of prior work has explored the limitations or friction introduced by existing hate and harassment mitigations. With respect to community guidelines, Pater et al. explored how multiple platforms prohibit engaging in hate and harassment, but that the associated terms or rules provide potentially inconsistent, non-exhaustive criteria [53]. This can lead to frustration among targets reporting presumed-violative content when they later learn an attack is not within scope of remediation [6, 8]. Similarly, Crawford et al. raised concerns about the suitability of flags as a response mechanism in general [14], while Kou and Gui explored why members of the gaming community do not flag toxic behavior—for instance, due to normalization [34].

Ambiguity around what constitutes hate and harassment also impacts automated content moderation tools. Here, bias in underlying training data can result in models being overly sensitive to certain terms [18], while majority-voting leads to gaps in protection for some communities [24, 35]. Automation also carries a risk of alienating audiences—and community moderators—due to a lack of transparency around expected standards of behavior [31]. Conversely, manual review incurs an emotional burden on the part of moderators [19, 70]. Thomas et al. provide a more comprehensive systematization of the design directions currently under consideration by the security community to address hate and harassment that balance these competing requirements [64]. We build on these perspectives by engaging with creators—many who rely heavily on the aforementioned systems—to understand what gaps in protection they experience and how they would address any perceived limitations.

2.3 Estimates of hate and harassment online

Pew Research Center previously identified that 41% of U.S. adults have personally experienced hate and harassment online [54]. Threats considered in their survey included offensive name-calling, purposeful embarrassment, physical threats, stalking, and sexual harassment. Participants engaged in a variety of protective practices including blocking (49%), reporting (22%), and taking a step back from platforms (7%). A similar study by Data and Society found targets of harassment turned to reporting (27%) or disconnecting from devices or platforms (26%) [16]. Vitak et al. explored harassment experiences specifically among female students, including how threats differ across demographic factors and social media platforms [67]. Our study refines the set of attacks considered as well as captures frequency-based estimates of attacks (e.g., “sometimes”, “often”) rather than binary estimates. Where possible, we compare our results against prior attack estimates or coping practice breakdowns for general internet users to differentiate the at-risk experiences of the creators in our study.

2.4 Broader at-risk context

Over the past several years, the human-computer interaction and computer security communities have undertaken a growing body of research about the experiences and needs of people who are at a heightened risk of being targeted with digital privacy or security invasions. This includes studies of people who may be at greater risk because of *who they are* (e.g., transgender people [12, 37, 58], children [74], older adults [23, 30, 50, 51], undocumented immigrants [25], refugees [60], people with visual impairments [2, 3, 28], people with learning disabilities [44], people experiencing financial instability [61, 68]); *where they are* (e.g., people in Bangladesh [1]); *who they are with* (e.g., survivors of intimate partner violence [22, 27, 46, 65, 66, 76], others being targeted by an intimate relation [38], survivors of trafficking [11]); or *what they do* (e.g., activists [15, 63], journalists [48, 49], sex workers [7, 47, 62], people involved with political campaigns [13]). Intersections between these attributes can further exacerbate risk (e.g., low-income African Americans in New York City [20], women in South Asia [56, 57]). According to the contextual factors laid out by Warford et al. [69], we consider creators to be at-risk due to their *prominence* as well as *access to a sensitive resource* (e.g., their audience).

Understanding the experiences and unmet needs of at-risk users—as we endeavor to in this study—enables the design of more equitable security and privacy protections for people who can be disproportionately harmed when their particular needs and contexts are not well understood. Across many of these studies we see that solutions designed only for the general population are insufficient for people who are at-risk, who may face increased harm or more frequent threats. The insights from work with at-risk users often leads technology designers to solutions that not only help at-risk users, but also greater protections for all technology users.

3 METHODOLOGY

We conducted a survey of creators from platforms including Facebook, Instagram, TikTok, Twitter, and YouTube on their experiences with hate and harassment. We outline key aspects of our survey design process, the demographics of respondents, the statistical analysis techniques we used, and the ethical procedures we followed to minimize harm.

3.1 Survey design

Our survey instrument consisted of four parts. We started with a consent form and then asked participants about their general attitudes towards being a content creator online, including the quality of their interactions with audience members. We then asked their level of concern towards experiencing hate and harassment as a creator due to their personal attributes (e.g., age, gender, race), personal beliefs (e.g., religious or political views), the types of content they post, their relationships (e.g., other creators, family, friends), or their increased visibility online, using a five-point Likert scale ranging from “Extremely concerned” to “Not at all concerned”. Given that previous research has shown there is no canonical definition of hate and harassment [53], we asked participants to consider a range of experiences including “bullying, threats, violence, unwanted sexual advances, stalking, toxic comments, and more.”

The second part of the survey asked participants about their personal experiences with 12 distinct categories of hate and harassment including bullying, trolling, or offensive comments; threats of violence; having chats, photos, or personal information leaked; and account hijacking. We selected these experiences from an existing taxonomy of online hate and harassment [64] and used either a five-point Likert scale ranging from “Always”, “Often”, “Sometimes”, “Rarely”, or “Never” for potentially frequent experiences (e.g., bullying), and a four-point Likert scale ranging from “Many times”, “A few times”, “Once or twice”, or “Never” for potentially infrequent experiences (e.g., account hijacking).[1] Associated with these experiences, we asked participants in aggregate to detail the number of platforms where they experienced such attacks, the types of aggressor(s) involved, and an open-ended section where they could share details about one such experience.

The third part of the survey investigated how participants responded to or coped with hate and harassment. This included their opinions of platform policies, reporting, and content moderation tools; the people and resources they turn to when experiencing harassment; and their thoughts on the largest gaps in existing platform-protections. The final part of the survey asked participants about advice and solutions that would help keep creators safer from hate and harassment. We concluded by asking participants to report their demographics, how long they have worked as a creator, which platforms they participate on, and the types of content they produce and their target audience. Our full survey instrument and consent form can be found in our Appendix.

3.2 Participant recruitment and demographics

We recruited creators over a period from June 24–August 14, 2021 from a residency program where creators opt-in to participating in research. Participation in this program is restricted to U.S.-based creators who are at least 18 years old. We first distributed our survey as a pilot to 10 creators to validate our instrument before expanding distribution to 198 creators. We received 145 responses. We removed 10 incomplete responses, yielding our final sample size of N=135. Respondents took a median of 15 minutes to complete the survey. The residency program compensated all respondents $15 USD for participating.

We report the demographics of our participants in Table 1. Of respondents, 56% identified as women, 40% as men, and 1% as non-binary. Another 1% preferred to self-describe, and 1% preferred not to say. Respondent ages fell into the following ranges: 18–24

[1] We based our determination of what attacks were frequent or infrequent based off previous estimates of prevalence among general internet users [16, 54, 64].

Demographic	Group	N	%
Gender	Woman	75	56%
	Man	54	40%
	Nonbinary	2	1%
	Prefer to self describe	2	1%
	Prefer not to say	2	1%
Transgender	No	132	98%
	Yes	0	0%
	Prefer not to say	3	2%
Sexuality	Heterosexual	102	76%
	LGBQ+	19	14%
	Prefer not to say	12	9%
	Prefer to self describe	2	1%
Age	18–24	22	16%
	25–34	37	27%
	35–44	40	30%
	45–54	23	17%
	55–74	12	9%
	Prefer not to say	1	1%
Race or Ethnicity	White alone	74	55%
	Black or African American alone	21	16%
	Hispanic or Latino	16	12%
	Asian alone	8	6%
	Two or more races	6	4%
	Prefer not to say	10	7%

Table 1: Demographics of the 135 creators who participated in our study.

(16%), 25–34 (27%), 35–44 (30%), 45–54 (17%), 55–74 (9%), while 1% preferred not to say. With respect to sexuality, 76% of respondents identified as heterosexual, while 14% identified as LGBQ+, 9% preferred not to say, and 1% preferred to self-describe. We asked respondents to provide their race or ethnicity as an open-ended response, which we manually coded into White alone (55%), Black or African American alone (16%), Hispanic or Latino (12%), Asian alone (6%), or Two or more races (4%). Another 7% of respondents did not disclose a race or ethnicity. We collected this sensitive, detailed demographic data in order to model the hate and harassment experiences of potentially higher risk demographic cohorts. We expand on the best practices that we followed when collecting such data later, in our ethics and anonymization processes.

Table 2 lists which platforms were used by respondents as part of their online presence. YouTube, Instagram, Facebook, Twitter, and TikTok ranked amongst the most popular[2]. We find that respondents curated audiences on multiple platforms simultaneously: 96% of respondents used two or more platforms, 82% used four or more platforms, and 21% seven or more platforms. This is similar to general internet users who also span their online presence across platforms [75]. In terms of career experience, 50% of the respondents in our study had been active for 6 or more years, 31% between 3–5 years, 18% between 1–2 years, and 1% under a year. In terms of their largest audience size, 44% of respondents had an audience of less than 10,000 people, 22% between 10,001-50,000 people, 9% between 50,001-100,000 people, 13% between 100,001-500,000 people, and 12% more than 500,001 people. For all subsequent analysis, we refer to respondents as creators.

Online community	N	%
YouTube	133	99%
Instagram	113	84%
Facebook	96	71%
Twitter	88	65%
TikTok	66	49%
Discord	58	43%
Reddit	39	29%
Pinterest	38	28%
Twitch	38	28%
Snapchat	22	16%
Others (please specify)	10	7%

Table 2: Platforms where creators who participated in our study create content.

3.3 Analysis methods

When comparing ordinal values from Likert survey responses, we first performed a rank-based omnibus statistical significance test using Kruskal-Wallis (KW). In the event of significance, we calculated the post-hoc Wilcoxon rank-sum statistic for pairwise samples to understand which pairs differ. We relied on the Python library `scipy.stats` for calculating all of these statistical tests.

Additionally, we modeled the outcome of experiencing each type of hate and harassment as a binomial distribution $Y_i \sim B(n_i, \pi_i)$ using a logarithmic link function. We binarized five-point Likert scales, treating the top three scales (e.g., "Always", "Often", or "Sometimes") as positive samples, while treating the bottom two scales (e.g., "Rarely" or "Never") as negative samples.[3] The model's parameters consist of categorical variables related to a creator's demographics (previously listed in Table 1) as well as the creator's audience size and years of experience. We omitted all samples where at least one demographic feature reported was "Prefer not to say," leaving N=116 samples for modeling. We relied on the Python library `statsmodels.api` for all modeling.

Throughout the paper we also include quotes from four open-ended questions from the survey where our participants described:

- A negative experience
- How they avoid hate & harassment
- Gaps in tools & coping practices
- Advice they would give to other creators

We provide these quotes to add color and context to the descriptive statistics and analyses described above, but did not perform a structured coding analysis of the open-ended responses.

[2] Other platforms specified by respondents included Patreon, LinkedIn, and Amazon.

[3] For four-point Likert scales associated with rarer events (e.g., account hijacking), we treated "Once or twice", "A few times", or "Many times" as a positive sample, and "Never" as negative samples.

3.4 Research ethics & anonymization

To ensure our work did not put participants at undue risk when recalling past sensitive experiences, our study plan was reviewed by a set of experts at Google in domains including ethics, human subjects research, policy, legal, security, privacy, and anti-abuse. We note that Google research does not require IRB approval, though we adhere to similarly strict standards. We alerted participants that our survey collected sensitive demographic data in our consent form. Additionally, all demographic questions included an option to "Prefer not to say", and did not require any answer within our survey tool.

We took multiple steps to ensure the anonymity of participants. We note that our survey instrument never collected names, email addresses, social media handles, or other public identifiers of creators. Distribution and compensation was handled solely by the organizers of the residency program who had access to identifying contact information, while the researchers involved in the study were the only parties with access to raw response data. Throughout the paper, the quotes we provide are the unedited responses of participants. We have only removed identifying information, including specific platform and community names or features to protect the participants from de-anonymization.

3.5 Limitations

Our recruitment strategy limits our participant pool to current, US-based creators. This excludes the possibility of engaging former creators who have since abandoned all their online presences due to hate and harassment. As such, the hate and harassment experiences of creators in our study may skew less severe. Similarly, we may be unable to capture any potential tipping point after which creators felt the risks of their career outweighed any other considerations. Our US focus also means we are likely to miss any cultural variations in the volume or severity of hate and harassment attacks in other countries. Lastly, our limited sample size of 135 creators means any statistical reporting—apart from more robust statistical tests that we perform—comes with a large confidence interval (8.43% for a 95% confidence level).

4 HATE & HARASSMENT EXPERIENCES

We explore the extent to which creators worry about or experience hate and harassment as part of their public online life. Through their personal stories and statistical modeling, we illustrate the heterogeneous threat landscape that affects all creators differently.

4.1 Hate and harassment concerns

Across all platforms, 71% of creators rated their interactions with other people as "somewhat positive" or "very positive", while 28% rated their interactions as "a mix of positive and negative". Only one creator rated their interactions as "somewhat negative", and none as "very negative". Despite generally positive attitudes, when prompted about their concern with the potential for hate and harassment, many creators expressed at least some concern as shown in Figure 1. The variations between these concerns are statistically significant (KW $\chi^2(5) = 23.2$, $p < .001$). We find that 51% of creators expressed moderate or higher concern with attacks "in general". Conversely, just 30% of creators expressed concern towards attacks "due to the types of content that they post" ($p < .001$). We note the variations between concerns around being targeted for personal beliefs, visibility, and personal attributes are not statistically significant.

In line with these concerns, many creators expressed that hate and harassment was unavoidable. As C-101 framed it: "*It's going to happen. Be prepared for it, expect it. Don't do anything stupid or shady online that would give them ammunition, then the worst they can do is call you names*". This was echoed by C-41: "*If you are on the internet, no matter the platform, get a thick skin. If you have issues dealing with people calling you names ... don't get on social media*". These concerns and attitudes shape how creators approach content creation and audience engagement.

4.2 Personal experiences with attacks

Over the course of their online careers, creators reported experiencing multiple forms of toxic comments and overloading (which involves amplifying toxic comments or negative ratings via coordinated action [64]) as shown in Figure 2. Overall, 72% of creators experienced at least one of the attacks "sometimes" or more frequently, while 30% experienced at least one the attacks "often" or "always". The variations between attacks are statistically significant (KW $\chi^2(5) = 124.1$, $p < .001$). Additionally, many attacks are not mutually exclusive—51% of creators reported experiencing at least two distinct types of attacks, and 37% at least three. We also asked creators whether they ever experienced potentially less frequent attacks—such as stalking, content leakage, or impersonation—as shown in Figure 3. Variations between this second set of experiences are statistically significant (KW $\chi^2(5) = 38.8$, $p < .001$).

Taken as a whole, 95% of creators reported experiencing some form of hate and harassment at least once during their career. By contrast, prior reports from Pew Research Center found 41% of U.S. adults have personally experienced hate and harassment online [54]. Similarly, Thomas et al. reported rates of 20–72% across 22 countries [64]. Our findings highlight the heightened hate and harassment experiences of creators compared to the general population. We discuss these attacks in detail below.

Bullying, trolling, and offensive comments. Bullying, trolling, and offensive comments—shortened as bullying—occurred more frequently then all other attacks in Figure 2, affecting 66% of creators at least "sometimes" (all $p < .001$). Only 10% of creators reported never experiencing bullying. This top ranking also mirrors the experiences of general internet audiences [16, 54, 64]. These attacks were noted as being both pervasive and frequent, as C-48 shared: "*I had one [audience member] during black lives matter movement who would send me harrasment [sic] videos and emails and messages every single day. I have [them] filtered to spam so [they] might still be doing it*". Creators, such as C-130, also highlighted that this risk was exacerbated during live events: "*I've had trolls come into the chat on live streams and disrupt it to the point of not being able to communicate*". C-34 felt this was due to the in-your-face nature of attacks while streaming: "*I think the key is that it's way more appealing to trolls to be able to harass in real time, such as in a chatroom or live stream, because they can get an instant reaction and they can see how many other viewers are there*".

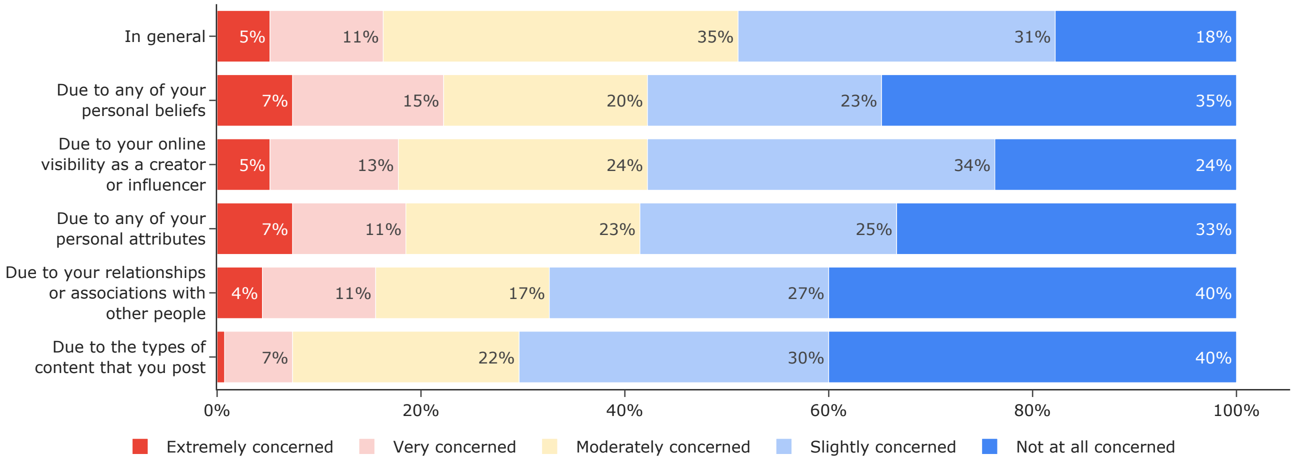

Figure 1: Creator concerns with being the target of hate and harassment, either in general or due to a specific factor, across all the platforms they participate on.

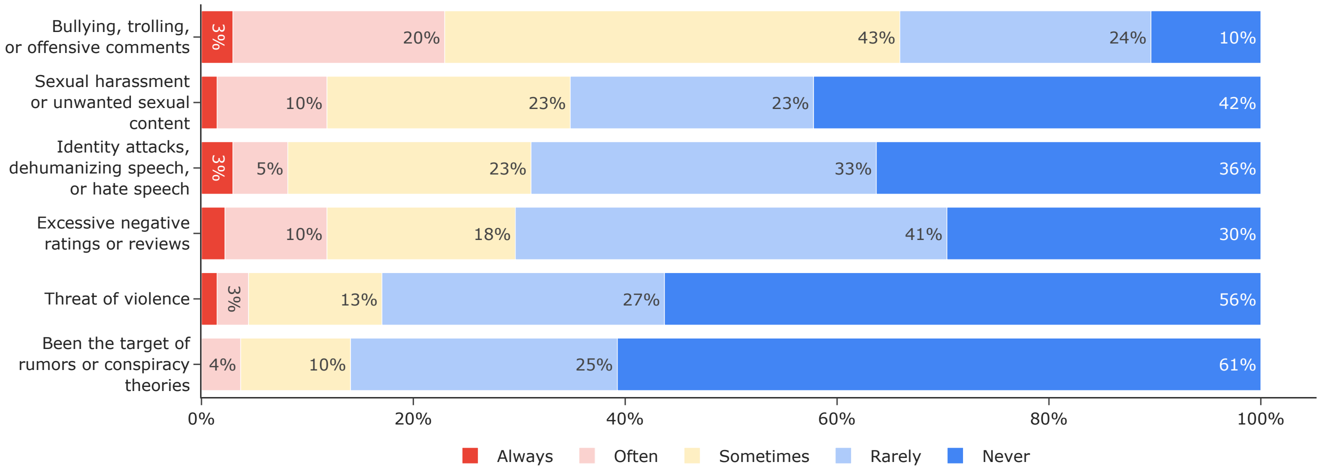

Figure 2: Creator experiences with toxic comments and overloading throughout their career. Bullying, trolling, and offensive comments were the most common experience, while rumors and conspiracy theories were the least common.

Sexual harassment, identity attacks, and negative reviews. Other prominent attacks that occurred at least sometimes included sexual harassment (35%), identity-related attacks (31%), and excessive negative reviews (30%). The variations between these are not statistically significant. When recalling an incident of sexual harassment, C-1 shared: "*On live streams we regularly get comments about taking off our cloths [sic]*". C-14 recalled both sexual harassment and identity attacks: "*This has happened mostly when I do live streams or sometimes when posting on [platform]. As a black creator I have been called racist slurs. As a woman I have had sexual content sent to me or comments made. ... It's common and a part of being a content creator. It shouldn't be, but it is*".

Threats of violence and rumors. Less prominent were threats of violence (17%) and rumors or conspiracy theories (14%). These were both statistically less frequent than all other attacks (all $p < .001$). While rarer, creators who experienced such threats often emphasized the severity, up to and including reaching out to law enforcement like C-33: "*Somebody didn't like [content] I posted and told me that he has mail [sic] a bomb to my house. Got FBI involved*". After trying to block a repeated offender, C-112 recalled how the attacks escalated: "*I had one guy that kept making fake accounts after I blocked him for repeatedly writing sexually inappropriate comments. He then sent me multiple death threats*". Similarly, C-44 shared: "*When an issues becomes extreme and you fear for your safety you can consult in [sic] local law enforcement but do not expect it to resolve the issue*". These experiences highlight how online threats can, at times, transform into offline, real-world physical safety risks.

Impersonation. 54% of creators experienced impersonation at least once or twice, ranking highest among all other attacks in Figure 3 (all $p < .001$). Creators discussed how attackers would impersonate audience members, or the creator directly, in order to sow discord. As C-79 recalled: "*Someone creates an account and bashes me, or they impersonate me and attack my community*". Similarly for C-52: "*someone or a group of people were impersonating my [audience members] (cloning) who were also on the live stream. Saying negative*

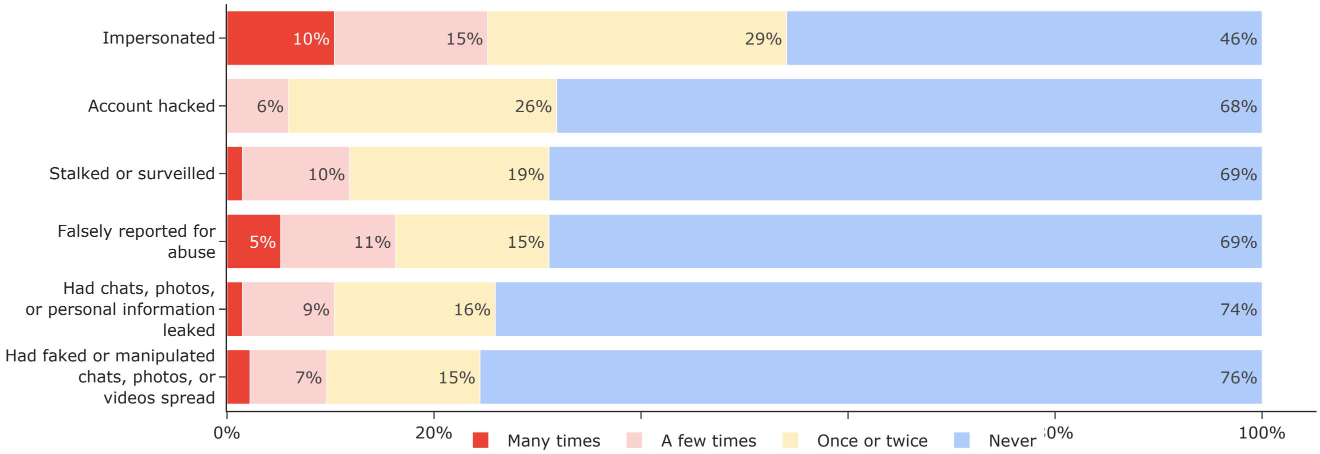

Figure 3: Creator experiences with other hate and harassment attacks throughout their career. Impersonation was the most frequent attack, while personal content leakage (such as "doxxing") and faked or manipulated media being spread were least frequent.

things". In other cases, creators suspected a financial incentive behind impersonation, such as C-82: "*A profile was created using my photos on other platforms such as [platform] and [another platform], and it was used to try and sell people goods/services*".

Account hijacking, stalking, false reporting, and more. Creators also recalled account hijacking (32%), stalking and surveillance (31%), false reporting (31%), having personal information leaked (26%), and having faked or manipulated content spread (24%) at least once or twice. We note there is no statistically significant difference between these remaining attacks. C-94 talked in detail about the threat of false reporting: "*someone that wants to harass you might not be posting comments, they can also falsely flag your videos. And the scary emails I've gotten from [platform] as a result were very nerve-wracking (not knowing if my content would be taken down or not)*". Reflecting on an incident of personal data being exposed, C-96 shared: "*A friend of mine's [messaging account] was allegedly infiltrated by someone who doesn't like the work I and a friend do on [platform] and malicious and almost certainly fake and out of context screen shots were shared to try and discredit us*". The audience interest around creators also means a range of attackers, either acting on parasocial relationships, such as C-69 experienced: "*I have a stalker. This person has sent a handwritten letter to my home address and has sent me multiple emails online*" to broad-based attacks as C-59 articulated: "*Had my information leaked and released to thousands. Also had my sites DDOS'ed. Only stopped after a few days of ignoring it*". All of these non-content-based attacks illustrate the potential limitations of content moderation alone.

4.3 Impact and reach of attacks

Apart from the frequency of attacks, we asked creators to reflect on the severity of the attacks, whether attacks reached across platforms, and other people targeted at the same time.

Severity of attacks. Thinking about all of their negative experiences, 68% of creators said they were "moderately", "slightly", or "not at all upset" by the aggregate attacks they faced, while another 25% stated they were "very" or "extremely upset". Many creators brought up being resilient to attacks, either brushing off incidents or moving on quickly. As C-41 shared: "*Deleted the comment. Hid user from the [community]. Over and done.*" C-55 expressed: "*I just delete those comments and, while they are hurtful, they don't have any lasting impact*". Even when creators are not initially upset though, the emotional toll of repeated attacks can add up, as C-75 shared: "*Of the multiple forms of harassment, the most common is in the form of hurtful comments directed at me or my family with no intent but to upset. These tend to be the least severe, yet are most common and ultimately impactful*". These reflections begin to show the nuance necessary to understand severity; most interactions are not individually harmful enough for a creator to fear for their safety or leave a platform, but the long term emotional toll of even seemingly mundane negativity can build up to real harm.

Platforms and attackers involved. For most creators, hate and harassment occurs across multiple platforms where they participate: 64% of creators mentioned attacks occurring on 2–3 platforms, while 8% recalled attacks on 4 or more platforms. The attackers are also not always random platform users, though 80% of creators recalled instances where that was the case. Instead, 42% of creators have been targeted by audience members, 23% by other creators, 20% by repeat offenders, and 17% by coordinated groups or online mobs. The attacker involved can affect how creators respond. For example, C-107 pointed out how they feel well-equipped to handle attacks that occur on their own content, but in other scenarios, they felt that there were no available controls: "*If I am harassed on my own [content] (like in the comments), then I feel that have the tools to handle the situation well. However, I'm not sure there's much I could do if I was harassed from another [creator] or in some other form that did not take place directly on my [content].*"

Collateral impact on others. As part of these attacks, creators mentioned the people affected went beyond just themselves. Indeed, 20% of creators recalled attacks that affected other creators simultaneously, while 15% stated attacks specifically targeted their audience, and another 7% that attackers targeted their friends or family. In the same way that one might experience multiple types of hate and harassment, the attacks may also purposefully have further reach. For example, C-62 experienced the direct impact of hijacking and impersonation, but the primary target was their audience: "*My account was hijacked and copied. Once they mimic it, they targeted my followers and asked them for money, pretending to be me. My audience reacted pretty quickly to tell me about it. I reported the incident, and it was resolved in 24-48 hours*".

4.4 Losing voices

While many creators brushed off attacks, we find evidence that some creators opted to self-censor or entirely leave platforms in response to attacks. By taking such actions, creators must balance the financial upsides of being public figures with the potential harms associated with hate and harassment. For 34% of creators in our study, the content they produced was either a major or primary source of income (over 50% of their income). For another 42% of creators, producing content was at least part of their income (more than 0%, but less than 50%).

Self-censorship. We asked creators whether they felt they could post the types of content they were most passionate about, without fear of retaliation or negativity. We found 22% of creators somewhat or strongly disagreed, while 8% neither agreed or disagreed. Hesitation around posting content in some cases turned into self-censorship. C-80 recalled hiding aspects of their identity in the wake of anti-Semitic attacks: "*[We] once released [content] in which we casually mentioned we were Jewish. Being Jewish was not a major part of the [content], it was a simple comment that took up maybe 10 seconds of the [content]. But we received hundreds of hate comments against Jews after that. It was so bad that we then blacklisted the words "Jew" "jewish" "judaism" etc so those words would no longer be allowed to be written in our [platform] comments section. I feel this was a good decision. Not only did we ban the word, but we decided not to mention our religion again in most future [content] for years after*". C-70 reflected: "*What I post and talk about is apolitical and does not touch on religion or other potentially sensitive topics. That is at least partially because I don't want to deal with the kind of people who seem to come out of the woodwork ... when people do post about those things. It kind of sucks to feel like I have to keep my opinions to myself*". These experiences illustrate the chilling effect that hate and harassment can have on creator speech online, particularly from at-risk communities who historically experience higher rates of attack.

Avoidance. Attacks also result in creators pulling back from platforms—including due to witnessing attacks on other creators. When asked whether they had ever left a community due to hate and harassment, 44% said yes, at least temporarily, and 19% said yes, permanently. For some, this stems from taking a moment to pause and reflect, as C-126 shared: "*[I] disconnect and get away for a short time. I've left [list of multiple platforms] for prolonged periods of time to get a handle on my emotions and take care of my life simply because people didn't like what I had to say*". For others though, it is a permanent decision to stay safer. As C-34 shared: "*Got doxxed by some disgruntled [audience members] in my old [community]. Fortunately I was able to work it out with them and contain it, but it's the main reason I decided to delete [platform] and not have a [community] on there anymore*". This sentiment was echoed by C-114: "*Demeaning language and mean-spirited behavior can quickly turn to violent threats of physical harm. It definitely curtails what, how much, where, and why I post*".

At the same time, creators who witness hate and harassment targeting others may opt never to participate with a community or platform, as C-34 shared: "*Seeing the harassment a lot of streamers deal with on [platform] is basically the #1 reason I don't plan to ever do live streams on [platform], it doesn't seem worth the trouble*". This sentiment was mirrored by C-19: "*In other online communities such as [platform], the culture there seems as though rudeness/harrassing speech comes with the territory so I choose not to engage on that platform very often as a result*". This can lead to a conscious choice of deciding where to participate, given the number of sites available as C-123 pointed out: "*There are so many platforms creators can be a part of, that I really only spend time on those which are most positive and stop uploading on the rest*". Our findings underscore the negative impact that hate and harassment can have on retaining creator voices online.

4.5 Factors that correlate with higher risk

We explored what factors—such as demographics, years of experience, or audience size—correlate with a higher risk of being targeted by attackers. We present the results of our modeling in Table 3 only for statistically significant factors ($p < .05$). We refer readers to our Appendix for complete model results.

Gender. Holding all other factors constant, creators who identified as women faced higher odds of experiencing rumors and conspiracy theories (20.0), sexual harassment (17.7), excessive negative reviews (4.5), and stalking and surveillance (4.0) compared to their counterparts who identified as men. This mirrors previous measurements of general internet users, where women reported higher rates of sexual harassment and stalking than men [54, 64]. Conversely, both creators who identified as women and those who identified as men reported statistically similar rates of bullying, identity attacks, threats, and other attacks. C-111 brought up how the experiences facing women are at times brought on by larger waves of hate: "*Over the decade we've used [platform] we have seen a number of more explosive, organized or targeted instances of intolerance and hate speech. A few times we've been targeted by malicious 'influencers' once by the user [redacted], another time by a Men's Rights online activist grouper over the so called '#gamergate'*".

Age. Creators older than 24 reported higher rates of multiple types of attacks. For creators between 25–34 years old, this included stalking (22.9), negative reviews (20.8), and identity attacks (9.1). Similarly, creators between 35-44 years old reported elevated rates of attack as enumerated in Table 3. We caution that the absence of statistically higher rates of attacks among 45–54 years old may be due to a limited sample size. These results differ from general population internet users, where 18–24 year olds experience the highest rate of hate and harassment [64]. As C-103 shared: "*Some people make negative comments because I am 60*". C-117, who identified as 45–54 years old, expressed a similar sentiment: "*I've been*

Factor	Attack type	Control	Treatment	Odds	P>|z|
Gender	Stalked or surveilled	Man	Nonbinary	111.3	0.018
	Been the target of rumors or conspiracy theories	Man	Woman	20.0	0.043
	Sexual harassment or unwanted sexual content	Man	Woman	17.7	0.000
	Excessive negative ratings or reviews	Man	Woman	4.5	0.022
	Stalked or surveilled	Man	Woman	4.0	0.026
Age	Excessive negative ratings or reviews	18–24	35–44	35.6	0.004
	Stalked or surveilled	18–24	25–34	22.9	0.005
	Excessive negative ratings or reviews	18–24	25–34	20.8	0.023
	Identity attacks, dehumanizing speech, or hate speech	18–24	35–44	14.2	0.007
	Bullying, trolling, or offensive comments	18–24	35–44	10.7	0.002
	Identity attacks, dehumanizing speech, or hate speech	18–24	25–34	9.1	0.032
	Account hacked	18–24	45–54	0.0	0.009
Race or Ethnicity	Bullying, trolling, or offensive comments	White	Non-white	0.2	0.012
Audience Size	Been the target of rumors or conspiracy theories	< 10K	1M+	232.9	0.002
	Been the target of rumors or conspiracy theories	< 10K	50-100K	58.9	0.020
	Threat of violence	< 10K	1M+	33.0	0.009
	Had chats, photos, or personal information leaked	< 10K	1M+	31.6	0.006
	Been the target of rumors or conspiracy theories	< 10K	100-500K	21.5	0.026
	Excessive negative ratings or reviews	< 10K	1M+	18.7	0.011
	Impersonated	< 10K	1M+	15.0	0.022
	Impersonated	< 10K	100-500K	13.5	0.001
	Threat of violence	< 10K	10-50K	9.8	0.017
	Account hacked	< 10K	1M+	7.6	0.037
	Excessive negative ratings or reviews	< 10K	50-100K	6.3	0.045
	Impersonated	< 10K	10-50K	3.2	0.040
Experience	Stalked or surveilled	1-2 years	6+ years	20.6	0.020
	Stalked or surveilled	1-2 years	3-5 years	18.2	0.024

Table 3: Odds of experiencing different hate and harassment attacks when holding all but one factor constant. Reporting limited to p < .05. Gender, age, race or ethnicity, audience size, and years of experience all correlate with varying levels of attack. Sexuality and transgender had no statistically significant correlations.

called names, been made fun of because of my age, and had my intelligence questioned. I try to ignore those comments, but sometimes feel the need to correct". C-103 expanded later: "*People in my age group are discriminated against on [platform]. We do not receive support*". One potential explanation is the mismatch between audience demographics—which skew younger—and older creators, which in turn can fuel attacks.

Race or ethnicity, transgender, sexuality. Our model does not detect statistically significant differences between LGBTQ+ and non-LGBTQ+ individuals, or between non-White creators and creators who identify as White alone. One exception is that non-White creators report lower odds of bullying and trolling (0.2). We caution the lack of differentiated risk may stem from our small sample size of LGBTQ+ and non-White creators, rather than the absence of a correlation.

Audience Size. Beyond demographics, a creator's audience size correlates with a higher likelihood of attack. For instance, creators with over a million followers faced higher odds of being targeted by rumours (232.9), threats of violence (33.0), having personal information leaked (31.6), and account hijacking (7.6) among other attacks. Creators with smaller, but still substantial audiences such as 100–500K followers, or 50-100K followers also exhibit similarly higher odds of attack as detailed in Table 3. Indeed, when we asked creators how hate and harassment changed after growing their audience size, 33% said that attacks had become a larger or much larger problem, while 22% said their situation improved, and 44% said it was the same. C-107 remarked on the potential risk that comes with increased exposure: "*Hate is universal and I feel like I've observed it directed towards every good creator I watch. It helps to know that even though it may seem personal, it isn't really and that the larger an audience you grow, the more likely you are to see it.*"

5 TOOLS, COPING PRACTICES, AND ADVICE

We find that creators rely on a variety of tools and coping strategies to contend with attacks. We explore each of these protective practices in detail, highlighting potential design opportunities, emergent advice for staying safer from threats, and the top recommended changes from creators.

5.1 Policies and platform-provided tools

We asked creators how effective policies and platform-provided tools—such as community guidelines, reporting, and content

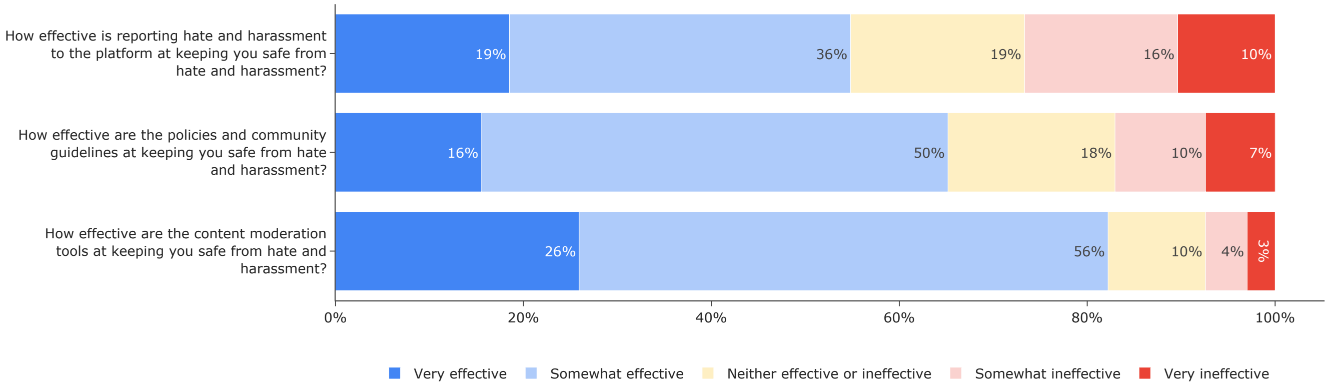

Figure 4: Creator attitudes towards platform-provided tools and policies at keeping them safer from hate and harassment. Reporting was seen as the least effective, while moderation was seen as the most effective.

Recommendation	N	%
Preventing anonymous or low-quality accounts from posting comments	75	56%
Automatic warnings to commenters that something they're about to post might be seen as hate and harassment	73	54%
Improved reporting tools to the platforms	68	50%
Automated moderation of comments	35	26%
Appointing community moderators	24	18%
Pinning positive comments	23	17%
Restricting interactions to only subscribers or followers	22	16%
Streamlined engagement with law enforcement	17	13%
Better external resources for mental health or bullying	16	12%
None of these would have much impact	5	4%

Table 4: Top changes recommended by creators for improving platform-provided tools and policies. Creators focused most of their suggestions on improving moderation, reporting, and preventing any circumvention of protections.

moderation—were at keeping them safe from attacks. Reporting was seen as the least effective and moderation tools as the most effective. Figure 4 shows the detailed breakdown of their responses (KW $\chi^2(2) = 20.0$, $p < .001$). We included an open-ended option to explain their reasoning and concerns, followed by an opportunity to select up to three top recommended improvements from a pre-defined list as shown in Table 4. We discuss each protection in detail below.

Content moderation. Moderation ranked highest among all of the solutions in terms of efficacy (all $p < .001$), with 82% of creators expressing that moderation was at least somewhat effective—and just 7% stating it was at least somewhat ineffective. When asked what gaps currently exist with moderation, creators expressed concern around attackers circumventing detection, the lack of default filters, the emotional burden of moderation, and the reach of tools. As C-80 shared: "*Sometimes when you block or ban a user, they are able to start a new account to harass you with*". Likewise, C-96 pointed to brittle implementations of filters: "*Sometimes racial slurs make it into the comments despite my having them listed as filtered terms requiring moderation*". To account for this, C-80 mentioned manually adding misspellings: "*3 words "ugly, annoying, and voice" block out about 80% of our hate comments! In total we have over 100 words blacklisted from our channel comments (account for misspellings as well!)*" However, establishing all of these filters is currently a manual process, with each creator duplicating the same keyword lists as C-32 shared: "*We shouldn't need to enter our own set of profanity filters for the words, though. The words should be part of an unacceptable speech filter*".

There is also an emotional toll to reviewing the correctness of automated decisions or manually removing abusive content, as C-40 shared: "*The worst thing about harassment is I have to see it. I can't afford someone to pre-moderate all my comments. I can remove it sure, but that doesn't help ME because I've already seen it*". Even when platforms allow creators to appoint community moderators, creators expressed there were too few available. As C-27 summarized: "*Too many home fires burning and not enough mods*". The burden of moderation can also resurface even for automated systems. C-57 pointed out that they "*get notified of negative comments even if [platform] deletes them. I stopped allowing email notifications because every time I got a negative comment, I was still able to see the comment, but once I tried to find it on [platform], it would be automatically deleted*".

Creators also expressed doubts on the reach of moderation tools, even when attacks were isolated to a single platform—and the possibility of overreach. C-127 pointed out: "*Content moderation tools help if the problem is on my [community] but not on other* *[communities on the same platform]. I can delete and silence haters on my [community] but they sometimes will go to other [communities] of the same type as mine and compliment that person to make sure the comment stays while being critical of me*". For some creators though, there is a risk of moderation going too far, as C-15 expressed: "*I think it is very important that [platforms] protects all users and creators from harassment and bullying but I also think that they have to be careful policing comments and content that is just a difference of opinion or disagreement*".

Taken as a whole, we find many creators strongly advocate for moderation, but they are conscious of current design limitations and the effort necessary to manually review abusive content. In terms of top recommendations, 56% of creators advocated for restricting posting from "anonymous or low-quality accounts", 54% for nudges to warn "commenters that something they're about to post might be seen as hate and harassment", 26% for better "automated moderation of comments", and 18% for "appointing community moderators".

Policies and community guidelines. 65% of creators felt platform policies and community guidelines were at least somewhat effective at keeping them safe from hate and harassment (ranking second, $p < .001$). Only two creators focused on policies when listing their concerns with existing protections. The few issues raised related to perceptions around selective application of community guidelines and being conscious of the limitations of any policy. On the theme of unfairness, C-89 pointed out: "*if a large massive content creator breaks rules and causes harm there rarely any repercussions that will follow that creator. They might be removed from [monetization, recommendations] but only for a very short time period and then they are back online like nothing has happened.*" Another concern was that while policies can be effective at times, it is true only up to a point. C-44 elaborated: "*There are times where policies no longer make a difference. In most cases we have the ability to "mute" "ban" or "report" people but in extreme cases there is no policy that will protect you.*" Our results suggest that policies and community guidelines are not top of mind for these creators. Instead, they prefer to focus on more practical actions that they are empowered to take.

Reporting. Just 55% of creators felt reporting was somewhat effective (ranking last, $p < .001$)—with 26% of creators negatively expressing that reporting was at least somewhat ineffective. Delving more deeply into reporting, we asked creators how often they reported attacks if they occurred. Just 41% of creators said often or always, 18% said sometimes, 36% stated they rarely or never reported, and 4% said they never report because they haven't experienced attacks. When discussing gaps related to reporting, multiple creators cited a lack of transparency or follow through, while some creators expressed friction in the process and uncertainty about what to report. On the theme of transparency, C-29 shared: "*I feel like when I report things it isn't taken seriously. I have no idea what the follow up is.*" Similarly, C-23 requested: "*I think there should be more transparency on how my reports are going*", while C-38 expanded: "*There is no known resolution to reports. Once we submit a report, it goes off into the ether and I have zero visibility into if there were any actions taken.*" Similar concerns around the value of reporting were also previously raised by general internet users experiencing hate and harassment [8].

Creators also mentioned limitations with reporting interfaces or having to make tradeoffs between reporting and other moderation actions. C-123 pointed out: "*I just block people but would love to have a [combined] option to "Report" the person for leaving an inappropriate comment. But I've been a creator for 10 years and have no idea how to do that? It should be an option 'delete, block, report hate' etc.*". Similarly, C-17 shared: "*If I remove the comment I don't have the option to report it. Many times I just want it down - I fear that by reporting it I have to leave it up until somebody looks at it.*" Confusion and friction around reporting flows thus reduces their use among creators.

As for what to report, C-90 advised creators to "*report everything*", while C-96 advised creators to "*report only serious offenses*", and C-34 stated "*I only ever bother reporting spam and scam comments*". These variations likely stem from the perceived lack of follow through on reporting hate and harassment, with some creators opting out of reporting entirely, or others focusing on categories of abuse (e.g., bots, automation) that they felt platforms responded to more seriously.

Taken as a whole, estimates of hate and harassment based on reporting volume alone may chronically under count the frequency of attacks today. At the same time, a critical channel is potentially missing for creators to seek help from platforms when they experience hate and harassment. In terms of the top changes requested by creators, 50% advocated for "improved reporting tools to the platforms".

5.2 Coping Practices

We also asked creators who they lean on for help and what resources, if any, they look to for advice when experiencing attacks. In responding to these questions, a theme emerged from many creators that hate and harassment was simply part of participating online. We explore each of their coping practices below.

Reaching out to others for help. 56% of creators said they at least somewhat agreed they had access to supporters who could help them cope with any stress or hurt stemming from hate and harassment. Comparatively, 21% of creators somewhat or strongly disagreed. We asked creators specifically who they would reach out to for help (Table 5). Creators commonly mentioned friends and family (70%), other creators (44%), and audience members (33%). As C-124 shared: "*Have a good support system of real friends that are not online, have support from family members and other creators but also be warned other creators are sometimes not your friends! This online world is very hard to navigate*". Similarly, C-134 suggested: "*Content creators should try to befriend other content creators who will better understand the type of negativity they may receive*". However, the systems to build these relationships are not always present and rely on creators organizing independently. C-129 requested streamlining this process: "*There should be a way to connect more creators together to build stronger and better communities to promote more unity and less hate*". In terms of top requested changes, 12% of creators favored better external resources for well-being.

Law enforcement was rarely considered an option (12%). As C-132 expressed: "*The problem is there are very few laws against this type of stuff and people know they can get away with anything*". C-134 suggested limiting such engagements to certain types of attacks:

Supporter	N	%
Friends or family	94	70%
Other creators or influencers	60	44%
My audience (e.g., followers, fans, supporters)	45	33%
Platform where harassment is occurring	36	27%
Health care provider or therapist	21	16%
Manager or other partners	16	12%
Law enforcement	16	12%
No one	17	13%
Other	3	2%

Table 5: Supporters that creators would turn to for help coping with hate and harassment. Most creators cited friends and family, while law enforcement and well-being resources were infrequently cited.

"*I've never had a serious case of stalking or violent threats, but I would suggest anyone receiving those should be screenshotting everything in case law enforcement needs to get involved*". C-71 echoed this sentiment: "*Someone bullying you on [platform]? Block and report them. Someone stalking you in real life? Go to the police. It all depends on the severity of the situation*". When engaging with law enforcement, C-132 suggested to "*screenshot and document everything, and involve a lawyer and the police*". Of top requested changes, 13% requested "streamlined engagement with law enforcement".

Turning to step-by-step guides. 50% of creators strongly or somewhat agreed they had access to online guides that provide step-by-step instructions for how to respond to hate and harassment. Another 21% at least somewhat disagreed, while 29% neither agreed nor disagreed. For some creators, these resources were made available via the platforms. As C-23 shared: "*Review all the resources [platform] offers ... watch YouTube videos, Google stuff, etc*". However, C-10 expressed such guides were not enough to help them navigate attacks: "*when dealing with issues, [platform support teams] typically sends you links to guidelines. We need better human to human interaction.*" Similarly, beyond step-by-step instructions, some creators expressed a desire for guides on how to cope with harassment. Specifically, C-88 requested "*education or therapy of how to deal with it, and how NOT to be affected by it.*" While multiple external resources exist such as the OnlineSOS action center,[4] Pen America's Online Harassment Field Manual,[5] or the Consumer Reports Security Planner (formerly from Citizen Lab),[6] no creators explicitly mentioned these resources. The exception was C-70 who mentioned "*For serious online abuse, contact the CrashOveride helpline,*" referring to the CrashOverride Network.[7] Taken together, many creators expressed a need for additional resources in this space, as well as greater visibility for such resources.

5.3 Advice for Other Creators

As the final part of our study, we asked creators what advice, if any, they would give to other creators coping with hate and harassment, or if there were any resources they would recommend. As these are unprompted, open-ended suggestions, we opted not to conduct a statistical or thematic analysis. Instead, we merely highlight directions we feel underscore the current thought process of creators with respect to security and privacy.

Have a plan. Prepare in advance. With hate and harassment on social media platforms seen as a pervasive, unavoidable part of being an online creator, the creators in our study felt it was important to plan for this eventuality. C-14 proposed a blueprint: "*You have to create a plan in advance. You can either ignore them - thus taking away the attention they want. Or call them out on it and get your audience involved.*" C-101 suggested one potential plan for creators: "*If you have any embarrassing past like me, do your best to purge everything from the web and move on with a clean break*".

Have a support group. Another piece of advice that can be built out in advance is to identify a support group. C-69 described who provided them with support (and who they avoid): "*It's important to reach out to people who understand. I reach out to fellow creators who have experienced the same thing. I do not reach out to my audience because that just pleases the people who are harassing you*". The importance of a support group is highlighted by the urgency C-41 details: "*Talk to someone regularly. Don't wait. If you are having dark thoughts related to hate and harassment, do not get on social media and seek counseling. Psychologist, psychiatrist, counselor, group of real friends - anyone that can help you get over any hate or harassment*".

Don't engage. While there are creators who have received viral attention for reaching out to trolls, most creators advised not to engage with people generating hateful content. C-106 suggested this only provides more "fuel": "*Don't engage with a hateful person/troll. Engaging them only gives them more fuel and they love engagement. Ignore and they do eventually go away*", and C-20 emphasized this with an even stricter dictum: "*I DO NOT engage, I delete and ignore. Here is my advice: NEVER go on [a platform] and call out the haters ... That adds fuel to the fire*". C-31 took a more pragmatic view: "*There's too many people in this world to try to please. Not possible to do it, so ignore it all.*" Not all creators agreed, such as C-14, who felt engaging was an advanced form of response: "*You can make it funny, you can repost them and call them out. It all depends on your personality. But if you're just starting out I would say just ignore and block until you can brush it off*".

Moderation is crucial. As we mentioned above, moderation tools were seen as highly effective, and the creators reiterated this, with practical advice for other creators on how to best use them, or general reminders that they shouldn't be ignored. C-43 pointed out the well-being benefits to making sure that creators have moderation plans in place: "*[Have] a moderator or assistant review comments. Making [sic] sure to filter out words you don't want to see in your comments. Mental health is important, protect it*", while C-32 discussed how these filters are important for protecting a creator's own community: "*I would recommend creators or influencers to set moderation words [on platforms] to prevent their own viewers from seeing comments of that nature*".

Keep things private. Creators also spoke to the benefits of general privacy and security advice and the risks of sharing private information online. C-22 stated this in the extreme: "*Be incredibly*

[4] https://onlinesos.org/action-center
[5] https://onlineharassmentfieldmanual.pen.org/
[6] https://securityplanner.consumerreports.org/
[7] http://www.crashoverridenetwork.com/

careful and share no information about your real life on the internet whatsoever". C-11 simplified this advice: "*Keep your private data hidden*".

6 DISCUSSION

Addressing hate and harassment remains a core challenge for the human-computer interaction and security and privacy communities. We synthesize our findings to act as a compass towards future research and improvements.

Ecosystem-nature of threats. The reach of attacks—across platforms and targets—has significant design implications. While most solutions focus on labeling individual comments or incidents, this fails to capture the ecosystem-nature of threats. There are a variety of actors operating on multiple different surfaces that can result in direct harms, or much more subtly, longer-term harms, to the creator or others in their orbit (family, friends, their community, other creators). Additionally, multiple creators highlighted how live streaming can increase both the frequency and severity of attacks. The CHI community could invest in better understanding these attacks at the ecosystem level: bringing our mixed-methods and broader socio-technical viewpoint to understanding these layered harms and dynamics. Building robust solutions requires understanding the nuances between attackers and tactics, as well as when early intervention can help mitigate attacks before they escalate into threats of violence, stalking, or other severe sources of harm.

Improving platform-provided tools. Creators consistently expressed that moderation was an effective day-to-day tool for combating hate and harassment. Nevertheless, they pointed to several possible improvements including pre-configured rules and filters, stricter controls on who can post comments (e.g., to prevent automated bots, low quality accounts), better community moderation support, and more automated decision making to reduce the emotional burden of manually reviewing abusive content. Creators were more skeptical towards reporting due to a perceived lack of follow-through. Examples of solutions currently proposed by researchers include Squadbox [42], which allows a person to appoint family members, friends, or community members to assist with review. With respect to stronger shared protections, practitioners have proposed curated keyword lists like Hatebase [26] and community-generated blocklists of abusive accounts like Blocktogether [9]. Equally important, prior studies have shown that audiences expect creators to be the ones to moderate and enforce rules on discussions [10]. Such enforcement needs to occur across all the platforms where creators have a presence, meaning a single platform's improvements to tools are not enough to fully protect a creator. Apart from requests for consistent tooling, our findings also illustrate that not all attacks that creators face come from comments or media posted to platforms. Account hijacking, stalking, false reporting, and having personal information leaked all operate outside the reach of moderation, necessitating a broader tool kit for creators—and general internet users—for responding to threats.

Better guides and well-being resources. Just 56% of creators in our study agreed they had access to support to help cope with attacks, and just 50% felt they were well-equipped with step-by-step instructions for how to mitigate or recover from attacks. In order to address this gap, creators articulated the need to streamline connecting with one another in order to share tips or simply to have an empathetic ear from someone who has experienced the same types of attacks before. Examples of such community building include HeartMob, which connects targets of harassment with supporters [8]. Apart from well-being and community, the advice that creators shared—while capturing a variety of security and privacy guidance—reflects a relatively nascent folk model for how to address hate and harassment. Curated advice targeting individual hate and harassment attacks (e.g., toxic comments vs. negative reviews) would better prepare creators for how to respond to threats when they occur, as well as how to better mitigate attacks in order to reduce their frequency or severity.

Lasting consequences of attacks. The fatalistic attitudes expressed by creators that attacks are simply part of online life underscores the importance of addressing hate and harassment. Otherwise we allow an increasing perception that social media is so toxic that many will choose to not participate. Our study presents evidence that today's creators currently engage in protective practices that curtail their own speech, such as engaging in self-censorship or quitting platforms entirely, to defend against the expected eventuality of experiencing online harms. And while these practices may greatly benefit individuals who would otherwise be targeted, they limit the richness of online communities. New protections—taken with an ecosystem-level approach—can thus shift this pervasive attitude and empower more creators to have truer voices online, particularly the diverse voices of at-risk populations most frequently targeted by attacks.

Beyond empowering creators. Creators in our study focused their recommendations on additional moderation tools and resources that would assist them in staying safer. This empowerment approach misses opportunities for technology designers and researchers to also consider more holistic solutions that automatically reduce exposure to hate and harassment—either via machine learning or conscious design—as well as to consider how to reduce the attacker's incentives behind engaging in hate and harassment. While these considerations were absent from our survey findings, we highlight these directions to ensure staying safer from hate and harassment is not perceived as solely the responsibility of creators.

7 CONCLUSION

We presented the results of a survey of 135 content creators that delved into their personal experiences with a variety of hate and harassment attacks; the protective practices they employ; and directional change they felt would better prevent, mitigate, or resolve attacks. Nearly every creator in our study experienced some form of hate and harassment, and for one in three, such experiences were a regular occurrence. As such, creators represent a population at-risk of hate and harassment compared to general internet users. Creators frequently relied on content moderation, and to a lesser extent reporting, for responding to attacks. However, when these platform-provided tools fell short, some creators engaged in protective practices such as self-censoring their personal attributes and beliefs, or leaving platforms and communities entirely, in order to avoid further harm. It is our contention that by understanding the personal stories of creators and the gaps they perceive with existing solutions, we can create a path towards reducing hate and harassment online for everyone.

REFERENCES

[1] Syed Ishtiaque Ahmed, Md. Romael Haque, Irtaza Haider, Jay Chen, and Nicola Dell. 2019. "Everyone Has Some Personal Stuff": Designing to Support Digital Privacy with Shared Mobile Phone Use in Bangladesh. In *The 2019 CHI Conference on Human Factors in Computing Systems.*

[2] Tousif Ahmed, Roberto Hoyle, Kay Connelly, David Crandall, and Apu Kapadia. 2019. Privacy Concerns and Behaviors of People with Visual Impairments. In *CHI '15: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.*

[3] Taslima Akter, Bryan Dosono, Tousif Ahmed, Apu Kapadia, and Bryan Semaan. 2020. "I am uncomfortable sharing what I can't see": Privacy Concerns of the Visually Impaired with Camera Based Assistive Applications. In *Proceedings of the 29th USENIX Security Symposium.*

[4] Julia Alexander. 2019. YouTube drafting 'creator-on-creator harassment' rules after Steven Crowder incident. https://www.theverge.com/2019/7/11/20691123/youtube-creator-harassment-vidcon-carlos-maza-steven-crowder-neal-mohan.

[5] Dennys Antonialli. 2019. Drag Queen vs. David Duke: Whose Tweets Are More 'Toxic'? https://www.wired.com/story/drag-queens-vs-far-right-toxic-tweets/.

[6] Zahra Ashktorab and Jessica Vitak. 2016. Designing cyberbullying mitigation and prevention solutions through participatory design with teenagers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.*

[7] Catherine Barwulor, Allison McDonald, Eszter Hargittai, and Elissa M. Redmiles. 2021. "Disadvantaged in the American-dominated Internet": Sex, Work, and Technology. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.*

[8] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from heartmob. In *Proceedings of the ACM on Human-Computer Interaction.*

[9] Block Together. 2019. A web app intended to help cope with harassment and abuse on Twitter. https://blocktogether.org/.

[10] Jie Cai and Donghee Yvette Wohn. 2019. What are Effective Strategies of Handling Harassment on Twitch? Users' Perspectives. In *Conference companion publication of the 2019 on computer supported cooperative work and social computing.* 166–170.

[11] Christine Chen, Nicola Dell, and Franziska Roesner. 2019. Computer Security and Privacy in the Interactions Between Victim Service Providers and Human Trafficking Survivors. In *Proceedings of the 28th USENIX Security Symposium.*

[12] Tya Chuanromanee and Ronald Metoyer. 2021. Transgender People's Technology Needs to Support Health and Transition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–13.

[13] Sunny Consolvo, Patrick Gage Kelley, Tara Matthews, Kurt Thomas, Lee Dunn, and Elie Bursztein. 2021. "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 30th USENIX Security Symposium.*

[14] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18, 3 (2016), 410–428.

[15] Alaa Daffalla, Lucy Simko, Tadayoshi Kohno, and Alexandru G Bardas. 2021. Defensive Technology Use by Political Activists During the Sudanese Revolution. In *2021 IEEE Symposium on Security and Privacy. IEEE Computer Society Press.*

[16] Data & Society. 2016. ONLINE HARASSMENT, DIGITAL ABUSE, AND CYBERSTALKING IN AMERICA. https://datasociety.net/output/online-harassment-digital-abuse-cyberstalking/.

[17] Discord. 2021. Developing server rules. https://discord.com/moderation/1500000176081-203:-Developing-Server-Rules.

[18] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society.*

[19] Bryan Dosono and Bryan Semaan. 2019. Moderation practices as emotional labor in sustaining online communities: The case of AAPI identity work on Reddit. In *Proceedings of the 2019 CHI conference on human factors in computing systems.* 1–13.

[20] Ame Elliott and Sara Brody. 2015. *Straight Talk: New Yorkers on Mobile Messaging and Implications for Privacy.* Technical Report. Simply Sedure. https://simplysecure.org/resources/techreports/NYC15-MobMsg.pdf

[21] Facebook. 2019. Community Standards Enforcement Report. https://transparency.facebook.com/community-standards-enforcement.

[22] Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. 2017. Digital Technologies and Intimate Partner Violence: A Qualitative Analysis with Multiple Stakeholders. *PACM: Human-Computer Interaction: Computer-Supported Cooperative Work and Social Computing (CSCW)* Vol. 1, No. 2 (2017), Article 46.

[23] Alisa Frik, Leysan Nurgalieva, Julia Bernd, Joyce Lee, Florian Schaub, and Serge Egelman. 2019. Privacy and Security Threat Models and Mitigation Strategies of Older Adults. In *Proceedings of the Fifteenth Symposium on Usable Privacy and Security.*

[24] Mitchell L Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 2021. The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality. In *ACM CHI Conferences on Human Factors in Computing Systems.*

[25] Tamy Guberek, Allison McDonald, Sylvia Simioni, Abraham H. Mhaidli, Kentaro Toyama, and Florian Schaub. 2018. Keeping a Low Profile?: Technology, Risk and Privacy among Undocumented Immigrants. In *2018 CHI Conference on Human Factors in Computing Systems.* ACM.

[26] Hatebase. 2019. The world's largest structured repository of regionalized, multilingual hate speech. https://hatebase.org/.

[27] Sam Havron, Diana Freed, Rahul Chatterjee, Damon McCoy, Nicola Dell, and Thomas Ristenpart. 2019. Clinical computer security for victims of intimate partner violence. In *Proceedings of the USENIX Security Symposium.*

[28] Jordan Hayes, Smirity Kaushik, Charlotte Emily Price, and Yang Wang. 2019. Cooperative Privacy and Security: Learning from People with Visual Impairments and Their Allies. In *Proceedings of the Fifteenth Symposium on Usable Privacy and Security.*

[29] Donald Hicks and David Gasca. 2019. A healthier Twitter: Progress and more to do. https://blog.twitter.com/en$_u$s/topics/company/2019/health-update.html.

[30] Rebecca Jeong and Sonia Chiasson. 2020. 'Lime', 'Open Lock', and 'Blocked': Children's Perception of Colors, Symbols, and Words in Cybersecurity Warnings. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.*

[31] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of Reddit Automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)* (2019).

[32] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. In *Proceedings of the ACM Transactions on Computer-Human Interaction.*

[33] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R Brubaker, and Casey Fiesler. 2019. Moderation challenges in voice-based online communities on discord. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23.

[34] Yubo Kou and Xinning Gui. 2021. Flag and Flaggability in Automated Moderation: The Case of Reporting Toxic Behavior in an Online Game Community. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.*

[35] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing Toxic Content Classification for a Diversity of Perspectives. In *Proceedings of the Symposium on Usable Privacy and Security.*

[36] Michael Lee. 2021. Twitch Streamer Nate Hill Swatted While Streaming Fortnite. https://gamerant.com/twitch-faze-nate-hill-swatted/.

[37] Ada Lerner, Helen Yuxun He, Anna Kawakami, Silvia Catherine Zeamer, and Roberto Hoyle. 2020. Privacy and Activism in the Transgender Community. In *2020 CHI Conference on Human Factors in Computing Systems.* ACM.

[38] Karen Levy and Bruce Schneier. 2020. Privacy threats in intimate relationships. *Journal of Cybersecurity* 6, 1 (2020).

[39] Taylor Lorenz. 2021. Facebook plans to pay creators $1 billion to use its products. https://www.nytimes.com/2021/07/14/technology/facebook-payments-creators.html.

[40] Taylor Lorenz. 2021. Young Creators Are Burning Out and Breaking Down. https://www.nytimes.com/2021/06/08/style/creator-burnout-social-media.html.

[41] Taylor Lorenz. 2021. YouTube begins paying out $100 million to creators using its short-form video feature. https://www.nytimes.com/live/2021/08/03/business/economy-stock-market-news/youtube-begins-paying-out-100-million-to-creators-using-its-short-form-video-feature.

[42] Kaitlin Mahar, David Karger, and Amy X. Zhang. 2018. Squadbox: A Tool To Combat Online Harassment Using Friendsourced Moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.*

[43] Enrico Mariconti, Guillermo Suarez-Tangil, Jeremy Blackburn, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Jordi Luque Serrano, and Gianluca Stringhini. 2019. "You Know What to Do": Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing.*

[44] Sonali Tukaram Marne, Mahdi Nasullah Al-Ameen, and Matthew Wright. 2017. Learning System-assigned Passwords: A Preliminary Study on the People with Learning Disabilities. In *Proceedings of the Thirteenth Symposium on Usable Privacy and Security.*

[45] Louise Matsakis. 2021. TikTok Is Paying Creators. Not All of Them Are Happy. https://www.wired.com/story/tiktok-creators-fund-revenue-sharing-complaints/.

[46] Tara Matthews, Kathleen O'Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F Churchill, and Sunny Consolvo. 2017. Stories from survivors: Privacy & security practices when coping with intimate partner abuse. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.* ACM, 2189–2201.

[47] Allison McDonald, Catherine Barwulor, Michelle L. Mazurek, Florian Schaub, and Elissa M. Redmiles. 2021. "It's stressful having all these phones": Investigating Sex Workers' Safety Goals, Risks, and Practices Online. In *Proceedings of the 30th USENIX Security Symposium.*

[48] Susan E. McGregor, Polina Charters, Tobin Holliday, and Franziska Roesner. 2015. Investigating the Computer Security Practices and Needs of Journalists. In *Proceedings of the 24th USENIX Security Symposium.*

[49] Susan E. McGregor, Franziska Roesner, and Kelly Caine. 2016. Individual versus Organizational Computer Security and Privacy Concerns in Journalism. In *Proceedings on Privacy Enhancing Technologies.*

[50] Andrew R. McNeill, Lynne Coventry, Jake Pywell, and Pam Briggs. 2017. Privacy Considerations when Designing Social Network Systems to Support Successful Ageing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems.*

[51] Helena M. Mentis, Galina Madjaroff, and Aaron K. Massey. 2019. Upside and Downside Risk in Online Security for Older Adults with Mild Cognitive Impairment. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.*

[52] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383.

[53] Jessica A Pater, Moon K Kim, Elizabeth D Mynatt, and Casey Fiesler. 2016. Characterizations of online harassment: Comparing policies across social media platforms. In *Proceedings of the 19th International Conference on Supporting Group Work.*

[54] PEW Research Center. 2017. Online harassment 2017. https://www.pewinternet.org/2017/07/11/online-harassment-2017/.

[55] Reddit. 2015. Subreddit rules now available for all subreddits. https://www.reddit.com/r/modnews/comments/42o2i0/moderators$_s$ubreddit$_r$ules$_n$ow$_a$vailable$_f$or$_a$ll/.

[56] Nithya Sambasivan, Amna Batool, Nova Ahmed, Tara Matthews, Kurt Thomas, Laura Sanely Gaytán-Lugo, David Nemer, Elie Bursztein, Elizabeth Churchill, and Sunny Consolvo. 2019. "They Don't Leave Us Alone Anywhere We Go": Gender and Digital Abuse in South Asia. In *Proceedings of the Conference on Human Factors in Computing Systems.*

[57] Nithya Sambasivan, Garen Checkley Checkley, Nova Ahmed, Amna Batool, David Nemer, Laura Sanely Gaytán-Lugo, Tara Matthews, Sunny Consolvo, and Elizabeth Churchill. 2018. "Privacy is not for me, it's for those rich women": Performative Privacy Practices on Mobile Phones by Women in South Asia. In *Fourteenth Symposium on Usable Privacy and Security.*

[58] Morgan Klaus Scheuerman, Stacy M. Branham, and Foad Hamidi. 2018. Safe Spaces and Safe Places: Unpacking Technology-Mediated Experiences of Safety and Harm with Transgender People. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018).

[59] Anna Silman. 2016. A Timeline of Leslie Jones's Horrific Online Abuse. https://www.thecut.com/2016/08/a-timeline-of-leslie-joness-horrific-online-abuse.html.

[60] Lucy Simko, Ada Lerner, Samia Ibtasam, Franziska Roesner, and Tadayoshi Kohno. 2018. Computer security and privacy for refugees in the United States. In *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 409–423.

[61] Manya Sleeper, Tara Matthews, Kathleen O'Leary, Anna Turner, Jill Palzkill Woelfer, Martin Shelton, Andrew Oplinger, Andreas Schou, and Sunny Consolvo. 2019. Tough Times at Transitional Homeless Shelters: Considering the Impact of Financial Insecurity on Digital Security and Privacy. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

[62] Angelika Strohmayer, Jenn Clamen, and Mary Laing. 2019. 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*.

[63] Borislav Tadic, Markus Rohde, Volker Wulf, and David Randall. 2016. ICT Use by Prominent Activists in Republika Srpska. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*.

[64] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini (Eds.). 2021. *SoK: Hate, Harassment, and the Changing Landscape of Online Abuse*.

[65] Emily Tseng, Rosanna Bellini, Nora McDonald, Matan Danos, Rachel Greenstadt, Damon McCoy, Nicola Dell, and Thomas Ristenpart. 2020. The Tools and Tactics Used in Intimate Partner Surveillance: An Analysis of Online Infidelity Forums. In *Proceedings of the 29th USENIX Security Symposium*.

[66] Emily Tseng, Diana Freed, Kristen Engel, Thomas Ristenpart, and Nicola Dell. 2021. A Digital Safety Dilemma: Analysis of Computer-Mediated Computer Security Interventions for Intimate Partner Violence During COVID-19. In *2021 CHI Conference on Human Factors in Computing Systems*. ACM.

[67] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying Women's Experiences With and Strategies for Mitigating Negative Effects of Online Harassment. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing*.

[68] Jessica Vitak, Yuting Liao, Mega Subramaniam, and Priya Kumar. 2018. 'I Knew It Was Too Good to Be True": The Challenges Economically Disadvantaged Internet Users Face in Assessing Trustworthiness, Avoiding Scams, and Developing Self-Efficacy Online. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–25.

[69] Noel Warford, Tara Matthews, Kaitlyn Yang, Omer Akgul, Sunny Consolvo, Patrick Gage Kelley, Nathan Malkin, Michelle L. Mazurek, Manya Sleeper, and Kurt Thomas. 2021. SoK: A Framework for Unifying At-Risk User Research. arXiv:2112.07047 [cs.CY]

[70] Donghee Yvette Wohn. 2019. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[71] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*.

[72] YouTube. 2021. Learn about comment settings. https://support.google.com/youtube/answer/9483359.

[73] YouTube. 2021. Moderate live chat. https://support.google.com/youtube/answer/9826490.

[74] Jun Zhao, Ge Wang, Carys Dally, Petr Slovak, Julian Edbrooke-Childs, Max Van Kleek, and Nigel Shadbolt. 2019. I make up a silly name' Understanding Children's Perception of Privacy Risks Online. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[75] Xuan Zhao, Cliff Lampe, and Nicole B Ellison. 2016. The social media ecology: User perceptions, strategies and challenges. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 89–100.

[76] Yixin Zou, Allison McDonald, Julia Narakornpichit, Nicola Dell, Thomas Ristenpart, Kevin Roundy, Florian Schaub, and Acar Tamersoy. 2021. The Role of Computer Security Customer Support in Helping Survivors of Intimate Partner Violence. In *Proceedings of the 30th USENIX Security Symposium*.

APPENDIX

Please see the Supplemental Material for our consent form, survey instrument, detailed statistical modeling results, and a full list of top recommendations from creators.