“The Report Button is Just for Decoration”: Moderation Practices and Needs of YouTube Creators

Victoria Zhong, Meghna Nair, Susan McGregor, Damon McCoy, and Rachel Greenstadt

Abstract. For content creators, whose careers rely on digital visibility, hate and harassment are an occupational hazard that impacts creators’ mental health, and in cases where online harassment spills offline, physical safety. We interviewed 19 YouTube creators on their experiences with and strategies used to combat hate and harassment, focusing on platform-provided tools, to understand their needs and identify areas for improvement. While participants *did* report offensive content, they did not find the platform’s reporting feature useful and felt they could not rely on it for remediation or support. Instead, they primarily used platform-provided moderation tools, social media hygiene practices, and other creators’ influence to manage the abuse they receive. Additionally, we found that harassment extended beyond the overt abuse perpetrated by bad actors and included seemingly innocuous interactions from their audience. Creators thus had to factor both external threats and intracommunity dynamics into their threat model. The persistence of these issues across years of research suggests that, absent changes in incentives or policy reforms, it is unlikely that platform improvements alone will meet user safety needs. We discuss how external factors contribute to these challenges or constrain solutions.

1 Introduction

Online harassment is a pervasive risk of having an online presence (Thomas et al. 2021; Samermit et al. 2023), especially for those who must maintain an active digital life for professional reasons (e.g., journalists, influencers). In fact, almost all *content creators* or *creators*—that is, users who create and share content on social media platforms for public consumption—report having experienced some form of harassment (Thomas et al. 2022).

doi:10.54501/jots.v3i2.266

Though many people still report a desire to pursue content creation as a career (Malinsky 2024), creators often report a sense of resignation, viewing online harassment as an inescapable part of having an online presence (Holton et al. 2023).

While content creators may come from any background, research indicates that certain groups—e.g., women, LGBTQ+—are more likely to be targets of online abuse as well as more severe forms of harassment (Keighley 2022; Nadim and Fladmoe 2021). Online abuse may be episodic or sustained (Take et al. 2024; Goyal, Park, and Vasserman 2022; Veale 2020), and can also create ripple effects on individuals' offline lives, either through offline threats such as stalking or swatting, or through negative mental health impacts that can sometimes have tragic outcomes (Dastagir 2021). In some instances, friends, family, or even unrelated third parties may suffer harm as a result of abuse that begins online (U.S. Attorney's Office 2019).

In their foundational effort to characterize the diverse attacks and harms experienced by those who post content online, Thomas et al. (2021) use "online abuse" as an umbrella term, an approach we adopt here. With this framing, online abuse has been an area of substantial research interest within the computer security, privacy, and human-computer interaction (HCI) communities, as well as a subject of popular reporting, for well over a decade. While this has produced a rich array of insights into attacks, harms, and mitigation approaches related to online abuse—including those outlined in Thomas et al. (2021)—the continuously evolving ecosystem of platform features, rules, legal contexts, and economic incentives has continued to make investigation of these concerns a research imperative.

The increasing proportion of young people who expect to pursue content creation as a career—coupled with the unpredictable and varied nature of platform governance approaches—makes investigating creators' approaches to preventing and responding to online abuse especially important. Prior research indicates that creators' risk assessment regarding online abuse is dependent not only on the type of content they create but also on the platform they use, and that different tools, affordances, and interfaces for reviewing, moderating, and reporting content can affect creators' ability to respond to harassment effectively (Samermit et al. 2023; Thomas et al. 2022; Gröber et al. 2024). Creators operating on multiple platforms also report that online abuse can move from one platform to another in cyclical waves (Nadler, Crain, and Donovan 2018), making the job of responding to even a single harassment campaign a recurring obligation.

Prior work on online abuse indicates that creators often find existing reporting systems to be ineffective or unclear (Reynolds and Hallinan 2024). While researchers have suggested design interventions to enhance moderation systems (Jiang et al. 2023), the implementation of such changes among platforms has been uneven, and each platform's affordances are unique. As such, creators must often develop the equivalent of "workaround" methods for managing online abuse (Lewis, Marwick, and Partin 2021; Mariconti et al. 2019) on specific platforms.

In this work, we specifically sought to elaborate existing insights about creator strategies for managing online abuse on YouTube through interviews with creators. Specifically, we sought to understand:

RQ1: What strategies do YouTube creators employ to attempt to prevent or mitigate online abuse?

RQ2: What impact, if any, do YouTube creators feel using YouTube's reporting feature can play in those efforts?

To answer these questions, we interviewed 19 YouTube creators in a range of content genres about their experiences responding to abuse, with a principal focus on how they used the tools provided by YouTube.

While our participants did describe using the "Report" button for abusive content, they did not find it an effective resource for either harm remediation or support. Instead, they relied largely on the platform's moderation tools, social media hygiene practices, and the influence of their fellow creators to combat and mitigate the abuse. Our findings also reveal sources of occupational hazard and administrative burden not identified in works like Thomas et al. (2021), particularly in the form of "needy follower overload," in which creators must manage the (sometimes unrealistic or personally harmful) attention demands of their own community.

While we believe these findings contribute both novelty and nuance to the research landscape on preventing and mitigating online abuse, we nonetheless observe substantial overlap in our core findings and those of prior researchers, as shown in Table 1. This observation prompts us to reflect more broadly on the aims, scope, and assumptions of online abuse research in the computer security, privacy, and HCI communities. In so doing, we identify fundamental sources of systemic risk (see Section 6.2) within the creator economy that individualize harms to creators while centralizing benefits to platform companies. These conditions, we argue, require developing new interdisciplinary research designs that engage content creation as a modern employment context in need of evidence-based policy reform.

2 Background and Related Work

2.1 Risk Factors of Online Abuse

While all creators face some risk of online harassment (Jhaver, Chan, and Bruckman 2018; Valenzuela-García et al. 2023), the positionality of a creator—the type of content they create, the sociopolitical context of where they reside, as well as their place in the intersection of these combined attributes—contributes to the level of risk creators face. Those belonging to a marginalized or vulnerable population (Goyal, Park, and Vasserman 2022)—e.g., women (Nadim and Fladmoe 2021; Posetti et al. 2020), LGBTQ+

folk (Uttarapong, Cai, and Wohn 2021; Scheuerman, Branham, and Hamidi 2018), or those who express political, ideological, or social views that oppose those of the user perpetrating harassment (Gröber et al. 2024; Goyal, Park, and Vasserman 2022)—often face added risk. The latter is especially prevalent when the target appears to be in agreement or aligned with "politically correct" or "woke" views (Take et al. 2024; Jhaver, Chan, and Bruckman 2018). Indeed, Take et al.'s work on Kiwifarms users finds that potential targets for a coordinated harassment campaign are chosen not only for their content, but also for their identity and ideological views (Take et al. 2024)—that is, users are targeted specifically because of their identity or beliefs.

2.2 Resultant Risk of Online Abuse

The most common type of harassment creators face generally falls under *toxic content*, which broadly encompasses hateful content directed toward them, ranging from insults to sexual harassment, trolling, etc.; this risk is universal to creators in general (Thomas et al. 2021; Thomas et al. 2022).

Just as positionality can affect the riskiness of content creation, it also affects the types of harassment frequently encountered. Marginalized creators are not only more likely to experience specific types of hate, but they also encounter these threats more frequently than their peers. For example, threats faced by creators who are disabled, international creators, and those who create adult content manifest differently from those of the average creator as well as to each other. Disabled creators who are also LGBTQ+ face significantly more ableist hate compared to non-LGBTQ+ disabled creators (Heung et al. 2024). Content creators in Pakistan, especially those engaged in activism or discussing sensitive topics (e.g., religion, politics, sexuality), face an elevated risk of physical harm and offline harassment (Gröber et al. 2024). Creators on OnlyFans, a platform where creators primarily produce intimate content, contend with the threat of being outed as a sex worker as well as nonconsensual dissemination of their intimate content (Soneji et al. 2024).

2.3 Navigating Risk and Impact on Creators

Content creators must balance their desire for freedom of expression, community building, and profitability with their emotional and physical safety needs. However, creators often do not start out creating content considering safety needs because they lack awareness of the risks of content creation, and only adopt protective practices after experiencing an incident (Samermit et al. 2023).

Despite creative expression being a primary motivator for content creators, creators often engage in self-censorship, avoiding topics that may cause controversy or backlash, to attempt to avoid negative interactions (Samermit et al. 2023; Thomas et al. 2022). Women facing online abuse not only engaged in self-censorship, but often withdrew from or made themselves less visible in online spaces (Valenzuela-García et al. 2023; Chadha

et al. 2020). In the case of female journalists, some even considered leaving journalism due to harassment (Posetti et al. 2020).

Creators suffer emotional distress not only from being exposed to hate, but also from managing hate and seeking recourse. Creators are often re-exposed to offending content while moderating and seeking remediation (Goyal, Park, and Vasserman 2022). Creators need to maintain a level of hypervigilance to prevent and respond to content leakage, including preemptively attempting to remove personal information on the internet, a laborious and Sisyphean task—even only considering data that originates from data brokers and not attackers (Take et al. 2022). Creators also need to predict whether any content they upload could pose future risk (Samermit et al. 2023). Seeking remediation is often by itself burdensome (Goyal, Park, and Vasserman 2022), but remediation may also not be possible; creators cite inadequate responsiveness from both platforms (Thomas et al. 2022; Samermit et al. 2023) and legal entities (Goyal, Park, and Vasserman 2022; Samermit et al. 2023). While some creators are financially able to delegate some tasks to hired staff, the onus is often on the creator to prevent and respond to harassment (Wohn and Freeman 2020). This excess burden, combined with the lack of resources on *how* to respond to harassment, can lead to burnout among creators (Goyal, Park, and Vasserman 2022; Uttarapong, Cai, and Wohn 2021). These emotional and operational costs often grow as their audience grows.

2.4 Shortcomings in Platform Moderation and Reporting

2.4.1 Limitations of automated moderation

While platforms themselves *do* somewhat moderate content, because many of the tools they use are automated, this moderation cannot account for context and as such is prone to misclassifying what is and is not inappropriate (Duarte, Llanso, and Loup 2018) and can be circumvented using creative means (T. Davidson et al. 2017). Classifiers are trained on corpora of existing hate speech (Gorwa, Binns, and Katzenbach 2020), which may miss novel forms of hate speech. Moreover, T. Davidson et al. (2017) found that not only were amateur coders unreliable at identifying abusive language, but how coders decided on what was hateful was unsurprisingly informed by their own biases of what constituted hate speech. While the annotators did find racist and homophobic language to be hateful, sexist language was only deemed offensive. While false negatives may expose vulnerable groups to seeing hateful content and perpetuate biases on what is deemed hate, false positives may be discriminatory against already marginalized groups by potentially censoring not only those in marginalized groups (Pastor 2024), but also conversations around those groups—implying it is wrong to mention those groups in any context. Indeed, (T. Davidson et al. 2017) also found that annotators may label based on the presence of words that could be offensive, without considering whether they were used positively or negatively. For example, sentences containing the words “gay” or “queer” were labeled as offensive even when used in a positive context (T. Davidson et al. 2017). Similarly, those posting about racism have had their posts taken down or

banned for referring to "white people" while allowing content that calls for exclusion of, or is derogatory toward, a group of people, due to technicalities such as targeting only a subgroup of a protected group or because the insult was not in "noun form" (Grassegger and Angwin 2017). As such, automated filters do not provide adequate recourse toward creators facing online abuse, and worse, may even censor or punish creators who speak out about experiencing online abuse due to their identity.

2.4.2 Tension between censorship and protection

Platforms often contend with balancing combating hate and harassment with allowing freedom of expression (Grassegger and Angwin 2017). While platforms seemingly provide transparency in the form of policy guidelines, these guidelines are often ambiguous and inconsistent, leading to user frustration when they discover that content they believed violated guidelines did not (Blackwell et al. 2017). Indeed, creators have expressed dissatisfaction with reporting abusive content to platforms due to both its ineffectiveness and the lack of transparency regarding decisions made in response to reports (Thomas et al. 2022; Samermit et al. 2023). Reporting harassment could even be harmful to the target of harassment themselves, as some reports require personal information that is then forwarded to the person or group responsible for the abuse, who may then weaponize such information to further harm the target (Take et al. 2024). This is the case for disputing a Digital Millennium Copyright Act (DMCA) takedown request, as the accused has to provide enough personally identifiable information for the claimant to be able to file a lawsuit should they choose to file a legal claim against the accused (Google, n.d.), which can be used maliciously against the creator.

While prior work identified gaps in protections in platform-mediated moderation for users in general, subgroups often have unique challenges and may be differently affected than the larger population as a whole. As such, we sought to identify the gaps in protections salient to creators. In addition, prior works (Valenzuela-García et al. 2023) often define harassment as behaviors intended to cause harm; our work explicitly challenges this stereotype, highlighting that the creators' threat model consists not only of threat actors who intend to cause harm, but also those who did not intend harm but in the creator's view did cause harm. Our work surfaces novel insights and themes surrounding harassment involving the complexities of handling "well-meaning" harassment, the cathartic role of even ineffective reporting, and the role creator networks play in responding to harassment.

2.4.3 Systemic factors

Individual stakeholders—creators and platforms—exist within a broader ecosystem that influences the decision-making of each stakeholder. These influences vary depending on where the stakeholder is situated within the larger system (Barile et al. 2012). While creators are affected by the constraints and affordances of how the platform is designed, as well as the platform's policies, platforms are often constrained by shareholder interests

(Zimmer 2021) and regulation (Margan 2025). As such, understanding a problem space requires understanding how external forces shape it.

One such factor is *administrative burden*, the idea that seemingly innocuous policy implementation incurs hidden costs to users: compliance, learning, and psychological (Moynihan, Herd, and Harvey 2015). Digitization can alleviate some of the burden by simplifying required tasks and making processes more intuitive, such as by utilizing nudges in UI design (Moynihan et al. 2022). However, it can also introduce new burdens or amplify existing ones. Though digitization can make services more accessible, it can also make them exclusionary, as participation requires both a degree of digital literacy and access to the internet. Transitioning to a digital medium without adequate infrastructure to support and maintain the service may also make the user experience more challenging, as users must contend with bugs and glitches. For example, Bhuiyan and Baniamin (2025) observed this effect when Bangladesh moved to online birth certificate registration from the previous physical applications. The proof-of-identity documents required proved exclusionary to those who were unable to obtain the extensive list of documents, affecting primarily those from a lower socioeconomic or marginalized background.

Administrative burdens can be unintended consequences of overlooking the potential impact of a policy on its userbase. Such burdens may also be intentionally introduced, e.g., by undermining legislation or limiting the usage of particular features or services (Peeters 2020). Though administrative burden is often discussed in the context of state governance, it can also apply to local levels of administrative relationships, such as between companies and employees or platforms and users. We discuss our findings in the context of the larger literature and how external constraints and incentives make it difficult to deliver meaningful solutions to these problems.

3 Methodology

3.1 Human Subjects Protection

Our Institutional Review Board (IRB) has reviewed and approved our study protocol, interview materials, consent, and recruitment procedures. The topic of the study was clearly stated in the recruitment study. Prior to the interview, we provided all participants with a consent notice via email, and during the interview, we asked participants to reconfirm their consent to participate in the study, as well as for us to record the interviews and retain their emails for further contact. Participants were allowed to decline to answer any question if they were not comfortable answering, and it would not affect their eligibility to receive the compensation. Interview participants were compensated with a $20 Amazon gift card after completing the interview.

3.2 Interview Design

The first author conducted all the interviews, and the interviews were conducted in a semi-structured manner. We conducted the interviews using a participant-led approach where participants are given epistemic authority and the researcher's role is primarily to facilitate reflection. The following outline of our interview structure is meant to serve as a guide for us to make sure we have hit upon all the questions we wanted answers to. As such, some participants may have been asked fewer questions if they had already answered the question as a response to a previous one, and more questions if we required some clarification to their response.

We started by asking participants to describe their online presence, including what kind of content they make and what platforms they post their content on. We then asked participants what hate and harassment looked like to them in order to understand the type of content that creators felt was hateful and to contextualize participants' later responses about strategies they employed when seeing content they deemed abusive.

In the first part of the interview, we asked participants whether they had experienced or witnessed harassment, and to recount specific instances if they were comfortable doing so. For each incident, we asked participants about the context surrounding the incident, how the incident impacted them, and how they created content.

In the second part of the interview, we asked participants about their practices addressing and possibly preventing online harassment in general, as well as whether they were aware of any recourse they had available when facing harassment on YouTube. Next we asked whether they were aware of YouTube's reporting feature and whether they had used it. We then asked participants what motivated them to use or not use the reporting feature, what they expected to happen, and what they hoped would happen when they reported content on YouTube. While we primarily asked about motivations for reporting, participants often also discussed in which situations they felt reporting would or would not be an appropriate course of action. For the participants who have used the reporting feature on YouTube, we asked about their experience using the feature, including difficulties they had and their satisfaction with the process of reporting. We also asked participants what outcomes came of the reports they made, if they knew, and whether they were satisfied with the outcome.

In closing the interview, we asked participants what tools or features they would like to see on platforms to help them feel safer, and what resources they would like to see to support other creators dealing with harassment.

3.3 Recruitment and Demographics

The first author recruited participants by directly reaching out to YouTube creators (YouTubers) via an email address they had made public on either their channel or social media. The first author also contacted the creators via the contact page on their website

or on Twitter/X. The list of YouTubers we contacted consists of creators the authors and their associates watch, Social Blade's[1] top 100 US/Canada, creators who are on Nebula,[2] and creators who are suggested by YouTube when viewing the aforementioned YouTubers' videos.

We specifically limited the creators we recruited to YouTubers based in the US or Canada who create content that has a "host," where the creator and their personality are part of the content. This includes content that revolves primarily around the creator, or content where the creator gives their opinion about other content (i.e., storytime, commentary, media analysis/recap). This excludes YouTubers who create content that can be separated from the creator, such as YouTubers who primarily make films or short stories (i.e., analog horror creators, musicians whose content is primarily their own music rather than talking about others' music, creators who make animations).

We had interested participants schedule their interviews through Calendly, an appointment scheduling software. When creators scheduled their interview, they were asked for their preferred pronouns and what country they resided in. We interviewed participants as they were scheduled. All interviews were conducted in March and April 2024.

We reached out to a total of 239 YouTubers across all different subscriber counts (1.1K–1.2M), with a median subscriber count of 134K subscribers. Of the YouTubers we reached out to, 30 responded, and 21 YouTubers agreed to participate. Of those who agreed to participate, we interviewed 19 YouTubers; two YouTubers did not show up. Of the participating creators, 16 creators were from the United States, and three creators were from Canada. Nine use he/him pronouns, five use she/her, two use they/them, one uses any/all, one uses none, and one did not respond. One participant also disclosed that he is transmasculine. Additional demographics can be found in the Appendix, Table A.1.

3.4 Interview Analysis

Each interview was recorded and transcribed by Zoom's live transcription feature with participant consent. These transcriptions were edited for accuracy by the second author. After each interview, the first author coded the interview using line-by-line coding using the Charmaz method of coding with gerunds[3] (Charmaz 2014). We chose to code with gerunds—actions—as opposed to topics because our focus is to understand our participants and their thought *processes* and behaviors around moderation, reporting, and harassment rather than the topic itself.

After doing the initial coding of all the interviews, the first author used axial coding and inductively categorized the individual codes into themes and relationships between themes. We identified challenges creators had faced as content creators with regard to handling hate and harassment, both in reporting content on YouTube and moderating their

1. Social Blade is a social analytics platform for tracking creators' metrics (e.g., subscriber/view counts).
2. Nebula is a creator-owned, subscription-based streaming platform, largely used by YouTube creators.
3. Gerunds are present participles that function as a noun, generally ending in "-ing."

online presence. We also identified the methods creators used to mitigate harassment beyond moderation practices. For example, the codes "reporting being ineffective" and "reporting as symbolic" would be grouped under the theme "Disillusionment with Reporting," and "not seeing offensive content removed [after reporting]" and "platform lacking feedback on reports" under "Reporting Outcomes." We note that these categories may not be mutually exclusive, and certain initial codes may fit under multiple thematic categories, i.e., "setting norms for comments" would fall under the themes of "Community as Motivation" as well as "Discouraging Hateful Comments" (which would fall under the broader category of "Mitigation Strategies"). Throughout this process, all the authors would meet and discuss codes, themes, and relationships to reach consensus.

4 Limitations

As our study is qualitative, our findings are not generalizable to all YouTube creators. Our participant sample is limited to English-speaking YouTube creators in the US and Canada who had been creating content in March and April 2024; as such, our findings cannot account for differences in the experiences and needs of YouTube creators in other countries and non-English-speaking YouTube creators. Moreover, because our sample only consists of current YouTube creators—those still creating content at the time the interview with them was conducted—the experiences and needs of our participants do not account for those who left content creation due to online harassment. In addition, though we did try to recruit more centrist and right-leaning YouTube creators, they did not respond. As such, our sample skews toward those who have left-leaning political views, and we are unable to capture differences in concerns of YouTube creators across the political spectrum. Our goal, however, is not to generalize but to gain insight into ways existing systems can be improved.

5 Results

In the following, we outline the primary threats participants reported commonly facing (Section 5.1), participants' attitudes toward YouTube's report feature (Sections 5.2 and 5.3, **RQ2**), and strategies and tools creators use to prevent and mitigate hate and harassment they face (Section 5.4, **RQ1**). Going forward, we use the term *creator* to specifically refer to YouTube creators.

Many of our findings overlap with prior work, but our focus on reporting also surfaced novel insights and themes. We summarize our findings and if/where they are also discussed in prior work in Table 1.

Table 1: **Summary of new and previously reported findings.**

Type	Findings
New	1. Audiences can and do unintentionally harass creators (creating implicit demands on a creator's time or attention) 2. Reporting content as a form of symbolic empowerment 3. Needing to network with larger creators to obtain recourse
Previously Reported	1. Reporting has little impact (Thomas et al. 2022; Gröber et al. 2024) 2. Friction while reporting (Gröber et al. 2024) 3. Steep learning curves from harassment experiences (Samermit et al. 2023) 4. Mitigation via social media hygiene and moderation tools (Samermit et al. 2023; Thomas et al. 2022; Soneji et al. 2024) 5. Ignoring hate as a protective practice (Samermit et al. 2023; Thomas et al. 2022; Soneji et al. 2024; Uttarapong, Cai, and Wohn 2021; Gröber et al. 2024) 6. Moderation is burdensome to creators (Uttarapong, Cai, and Wohn 2021) 7. Censorship vs. moderation, by both platform and creator (Heung et al. 2024)

5.1 Threat Landscape

Out of all of our participants, only one creator (I16) did not report experiencing or witnessing content they felt constituted harassment. However, every participant *did* report having received some form of hate.

The most common threat participants reported having direct experience of was toxic content. In fact, most participants we interviewed had only experienced hate in some form of toxic content.

While participants were concerned with content leakage, specifically doxxing, only one participant had been doxxed. Table 2 provides an overview of the types of hate participants reported either having directly experienced or being concerned about.

Two participants had online harassment turn offline. One participant was at the time facing a sustained harassment campaign, and a venue hosting the convention the creator was speaking at received threats, resulting in the creator needing extra security at the event. This participant had also experienced online impersonation as part of the harassment campaign.

Aside from instances where (1) an attacker shares the creator's content to another platform with the intent to ridicule, attack, or in any way cause harm to the creator, or (2) toxic content appears to be perpetrated by a group of attackers, participants were split as to whether toxic content by itself was harassment, with some viewing it only as hate. However, creators agreed that any unwanted actions, both negative and positive, that were persistent or violated their boundaries constituted harassment.

Table 2: **Types of hate faced by participants grouped by categories.** All subcategories except Deadnaming and Falsified Abuse Report/Flag were threats experienced by at least one participant.

Category	Subcategories
Toxic Content	General Insults*, Hate Speech*, Threatening Language*, Incitement, Sexual Harassment, Misinformation and Conspiracy Theories, Off-Topic Rants, Deadnaming†
Content Leakage	Doxxing, Leaked Profiles
Surveillance	Cyberstalking
Impersonation	Impersonated Profile
False Reporting	Falsified Abuse Report/Flag

* May be directed toward the creator, another person, or a person the creator mentions in their content.
† Though Thomas et al. (2021) categorize deadnaming under content leakage, participants viewed deadnaming as hate speech used to invalidate a trans person's identity rather than to reveal information.

5.1.1 Toxic content

Toxic content often came in the form of comments that ranged from mild insults to hate speech. Unsurprisingly, participants who were women, PoC, LGBTQ+, or Jewish, however, experienced hate speech directed at the group(s) they belonged to as the type of toxic content they mostly received. Interestingly, a few male creators who did not present as belonging to any marginalized group would receive antisemitic remarks targeted toward them, even though they are not Jewish. For example:

> *It was just a few months ago, there was somebody who left a comment on a video of mine saying I was a [antisemitic phrase], which is strange. I know he's being antisemitic, but I'm not Jewish, so I'm not really sure where he was going with that.* (I12)

Similarly, another creator (I13) recalled "In the past, they have used the F slur when I'm not gay myself."When asked why they felt the person used these specific insults against them, despite not being part of the group the insults target, both alluded to the insults being convenient to use. I12 specifically states:

> *If they dislike something you say, they get mad, and they just instantly go to attacking race or gender, or something like that.*

Interestingly, for creators who produced content related to science, the most common form of toxic content they received was misinformation or conspiracies posted under their content—sometimes unrelated or only tangentially related to the topic discussed in the creator's video. One of these creators also had antisemitic remarks on their videos; however, these remarks were not targeted at them, but at those mentioned in the video in a conspiratorial manner.

> *There's been some antisemitic stuff I've gotten on some videos. You almost have to be well-versed in this sort of thing to even realize that they're saying*

> *something. That's a dog whistle. ... They're not directed at me but just the topics in the videos. ... It's again the conspiratorial sort of thing.* (I06)

5.1.2 Boundary violations

Boundary violations include not only sharing or accessing information the creator did not wish shared or accessed (i.e., content leakage and cyberstalking) and persistent unwanted interactions, but also demands on the creator's time, interaction, or content. While content leakage and surveillance may be perpetrated by both malicious users and fans, demands on the creator are primarily perpetrated by users who were, at the time, fans of the creator. One participant (I17) recounts an instance where a supporter had asked them to critique their work over email and became irate when the creator did not respond quickly enough.

> *I had one patron on Patreon who I had agreed to look at some of his work. I told him "I'm extremely busy and I don't know exactly when I'll be able to get to your stuff. I will, but I hope you'll have a little patience with me." About three weeks later, I write the stuff and send [him] an email. He was pretty rude, and he told me to never contact him again.*

Another participant (I10) had an audience member demand that they remove content they had made where they had collaborated with a brand that garnered controversy at the time. When the creator did not respond, the audience member attempted to incite harassment toward the creator via a callout post.

> *I had worked with [brand], and they had hired me for a collaboration. I can't turn something like that down, because I have bills to pay. ... There was not a boycott for [brand] yet; I started working with them [before the boycott]. Somebody DMs me [about the brand's recent controversy]. I just ignored it; then, a day or two later, they had posted on their own stories, tagging me. They messaged me again and had said, "I've reached out to you multiple times, giving you plenty of time to take down your post. Do it." So I blocked them. After I blocked them, they made a post about me on their Instagram. [Participant reads caption that names them and accuses them of exploiting their audience by working with the brand].*

Three participants reported having received unwanted sexual attention. Two of those had an experience where a fan had persistently made unwanted advances toward them, despite them not responding. In the case of one of these participants, the fan attempted to meet him in person after he had disclosed to his audience that he would be attending an event. Interestingly, while one of the participants *did* express that they felt unsafe because of their harasser, neither felt that the perpetrators intended to make them feel unsafe.

5.2 Attitudes and Motivations Around Reporting

5.2.1 Reporting as symbolic empowerment

Despite a lack of faith in reporting, most participants stated that they sometimes still report content the creator deems sufficiently hateful, abusive, or harassing, as a way of feeling like they are taking action to protect themselves and their audience. In I14's words:

> *I use it just for the catharsis. It's about as impactful to me as stabbing a voodoo doll would be, [but] it just feels good. I can strike back in a way even if it's not going to have meaningful consequences.*

Indeed, participants felt that reporting was "all they can do" (I09) and "there was nothing else they can do" (I14). One participant (I04) called it a "fingers-crossed situation," going in with no expectations, but hoping some action gets taken.

5.2.2 Reporting and free speech

A few creators felt that reporting content was antithetical to free speech. I18 specifically felt that only comments that cause material harm warranted a report and felt that even filtering comments by keyword could be considered censorship. I17 likened reporting to "running to mommy because somebody online made me feel uncomfortable" and felt that platforms shouldn't "be in charge of making decisions about who should post and who shouldn't," especially as platform moderation is only punitive and does not "make changes to what kind of content people are encouraged/discouraged to make."

Though participants spoke about wanting to uphold free speech out of principle, they were also worried about garnering a reputation of being defensive or engaging in censorship, which may impact their growth.

> *I don't want to be notorious online for reporting or blocking. I don't want to ban someone who then gleefully goes to 4chan and talks about how I was so scared of their thoughts ... YouTube deletes comments and then I will get blamed for it. One of my videos got its comments disabled by YouTube without telling me they had done it. It feeds into my distrust about YouTube as the leasing body of comment systems.* (I17)

Another creator recalled an incident involving another creator receiving backlash for filtering certain words.

> *There was a big controversy around a YouTuber called [name]. She was clearly restricting comments on their videos that used words such as freebooting or stealing because that was what she was accused of, and she drew a lot of criticism—rightly—for that because it made her seem like an oligarch or a dictator in her own space.* (I18)

5.3 Reporting as an Unproductive Method for Combating Hate

Most participants said they felt that reporting offending content was not an effective way to curb hate and harassment.

> *The [report] button is just for decoration, it doesn't actually do anything.* (I12)

5.3.1 Not believing reporting makes any impact

The major complaint participants had was that they were not seeing any impact their reports had on the reported content due to either not receiving feedback on their reports or seeing that content they had reported that violated YouTube's guidelines remained up. In the words of I04: "it takes a lot for a YouTube report to be successful."

I14 recalled having seen content that clearly violated YouTube policy, but YouTube did not take action on the content.

> *She used a very racist term in her thumbnail. It took people on Twitter getting very mad about it and only then she took that video down herself. If literally using a slur in your thumbnail doesn't get YouTube to take it down, I don't know what will.*

I19 had instances where they reported threats made against them, and YouTube did not take any action.

> *I've had death threats and I've reported it; YouTube didn't do anything.*

Participants also felt that YouTube's moderation is incredibly opaque and policies aren't enforced consistently, leading to feelings of distrust in YouTube's moderation.

> *There isn't that level of transparency. You don't know if it's effective, if the account is suspended, for how long, or the consequences for it.* (I01)

> *These videos can be reported to YouTube, and it's up to them how they want to enforce their policies. Their policies are not consistently applied, and therefore certain accounts are able to get away with breaking these policies.* (I07)

Participants also brought up that because hateful content will often use *dogwhistles*, coded language that looks innocuous to those unfamiliar with the rhetoric, successful reports would require those reviewing the reports to be aware of those dogwhistles.

> *They're [comments] often not so flagrant about it; it's more dogwhistly, and you would have to be aware of the content. You'd have to understand that they don't literally mean this.* (I14)

Participants noted that even if users were banned, it would be easy to create a new account to continue their behavior. Moreover, reporting would not be useful for organized

harassment campaigns due to the number of users participating, as well as the fact that harassment is not limited to YouTube specifically, but exists on all platforms.

The few participants who believed the reporting feature on YouTube was an effective mechanism had seen some action taken when reporting content themselves or had seen content removed and attributed the removal to the reporting mechanism.

> *I think so [reporting is effective]. On YouTube, most of them [my reports] have been pretty successful.* (I08)

> *I went in to remove the comment, but it was already gone by the time I had read the notification; so I assumed my audience had just flagged that one on my behalf.* (I18)

5.3.2 Experiencing friction when reporting

Some participants expressed that reporting was not very straightforward. Participants mentioned that there were more steps when filing a report on YouTube compared to other platforms, as well as confusion about which category their reports fall into.

> *Sometimes there's not a category for the thing I want to report it as. If you have something that doesn't quite fall into the categories, then you're not really sure where to go.* (I15)

Another participant brought up experiencing a lot of friction attempting to report content on YouTube's mobile application. Specifically, hitting the report button would open up the reporting form in a browser and require the user to sign in again before they can report the content. A few participants also felt that YouTube changes the process of reporting too frequently, and one (I04) wasn't sure how to even report a user at times.

> *It's harder, once upon a time you just had a button. I had to Google it at one point; it was not immediately obvious where it was at the time. It wasn't giving very helpful search results; I found a lot of outdated stuff. ... It doesn't feel like a big enough deal to go through the process of reporting it.*

5.3.3 YouTube lacking incentives to act on reports

Participants attributed the lack of impact of reporting to their belief that YouTube's "business model is about keeping advertisers happy, not about keeping users or creators happy" and that "there's no incentive to 'do the right thing' " (I19). Indeed, participants felt that YouTube's moderation decisions are influenced by financial incentives and prioritize the needs of stakeholders that either generate income or may negatively impact profit for the platform. Participants commonly mentioned observing that YouTube will quickly penalize even minor violations involving copyrighted media, but hateful content will take longer to remove if it is removed at all.

> *They'll take down a video if it includes 30 seconds of a copyrighted song, but if a video contains actual examples of bad content, YouTube will be very slow to take it down.* (I14)

> *They'll happily take your channel down in a heartbeat if you put three seconds of a song that's copyrighted. If you threaten someone's life, then nothing happens.* (I19)

Interestingly, one participant (I09) recounted a time when their own content was reuploaded in whole by another user with only minor visual edits and the audio translated into another language. They reported the videos, but did not receive any remediation from YouTube. They only got the videos removed once they contacted the user who reuploaded their content and threatened legal action; only then did that user themselves take down the videos.

Participants also expressed feeling that YouTube showed favoritism to larger creators because they were "an asset too valuable to lose." In the words of I07:

> *The more stature you have on the platform, the gentler YouTube seems to be. You really have to break the rules very aggressively when you're at those upper echelons.*

As such, some participants mentioned feeling like they needed to know larger creators to have their issues taken seriously. One creator, however, notes that the larger creator could potentially be a threat themselves if the creator has a falling out with them or the larger creator has ulterior motives for helping the smaller creator.

5.4 Strategies Creators Use to Prevent and Mitigate Hate

Participants described several methods they used to combat and prevent hateful interactions. In general, these methods fall into the categories of moderation, where participants curate comments left on content they have posted, and selective disclosure, where participants describe curating what they disclose to their audience and their interactions with their audience.

5.4.1 Moderation as mitigation

The most common moderating method participants utilized was the comment filter feature available to YouTube creators. This feature places any comment that is picked up by the filter into a moderation queue where creators can review the comments and decide whether they want to allow the comments to be posted. While many of our participants relied primarily on the default automated filter to help pick out comments that may be offensive, some participants also used their own custom list of words or phrases they wanted to block either because the language or words being used in hateful comments are unique to them, or because of unique circumstances where comments mentioning

specific topics are likely to be hateful. For example, one participant (I08) described having to add the word "curry" to her list of filtered words due to it being used in racial epithets directed toward her.

> *I've updated the filters on each platform to catch as many of them as I can but it's hard, because now I have to ban the word curry because people will frequently be like: [racial insult using the word curry]. Curry isn't inherently bad, but because they string together words that, by themselves, are innocuous, but together, are kind of trash, I have to filter that out too.*

Participants also utilized the "hide from channel" feature against those who commented hateful messages on their videos. This feature acts as a "shadowban" such that if the creator hides a user from their channel, comments the hidden user attempts to leave will not appear to anyone but the hidden user. This was seen as a more effective alternative to reporting, as the offending user would not know they would need to create a sockpuppet.

> *I [hide from channel] because if they get reported, they get their thing removed—they're going to make six burners anyways. [In the case of hiding], they just scream into the ether, and I think it's funny because they think they're commenting, but I can't see it; nobody else sees it.* (I03)

While the primary function of moderation was to remove hateful and offensive content from their page, participants described how moderation was also effective as a way to discourage hateful behavior by setting expectations of conduct in their comment sections. That is, they felt that if they allowed hateful comments to remain unchallenged in their comment section, those inclined to post hateful comments would see it as a space where such comments are tolerated. Similarly, if their comment section is free from hateful or abusive rhetoric, or such comments are challenged by either the creator or those in the audience, it may discourage those who wish to post hateful or abusive comments. Many participants also noted that they often felt it was best not to engage with those leaving hateful remarks, as we will describe in Section 5.4.2.

This also applied to wanting to influence a broader creator culture, as I17 felt that by removing bigoted comments, it may set an example for other creators to remove comments because they disagreed with those views.

> *There may be a comment that says this person doesn't write good poetry because they're Jewish, and to me that doesn't necessarily cause any material harm. I think it's terrible and I don't agree, but in the same vein, I want to try and be an example of a decent creator. I don't want to stifle ideas just because I disagree with them, because otherwise content creators who are dissimilar from me, they might see how I respond to those comments and might do the same on their channels where they block comments that are actually more constructive or positive towards these conversations.*

Participants also saw moderation as a method to protect their audience from seeing and potentially being affected by hateful comments.

> *You could insult me, that's fine, but I'm not letting you bully my audience.* (I03)

> *[Hate] doesn't have a huge impact on me, [but] what gets to me though is that people who may not have as thick of skin as me are consuming my content. They come across these comments and it might impact them more; so that's where I start getting a bit protective of my viewership in my community.* (I05)

This is especially true of creators whose audience consists of marginalized groups.

> *A lot of Black people watch me, so I don't need other Black people seeing that.* (I11)

> *I want people, especially women and other minority people, to feel comfortable and to be able to leave a comment and not be harassed.* (I19)

5.4.2 Social media hygiene as prevention

Participants also alluded to using social media hygiene techniques to both prevent hate from occurring and mitigate existing hate from escalating. Most participants alluded to ignoring hate comments, as they believe any acknowledgment of the hate received would encourage the "haters" by eliciting a response. Indeed, creators alluded to believing that one motivation for those who leave hate was to provoke a response, and as such, any acknowledgment by the creator can be seen as "the hate getting to them."

> *If you try and address or complain about it publicly, that just encourages more people to keep going. It's just bullies getting a rise out of their victim.* (I12)

This sentiment, however, was not ubiquitous, as a few creators felt that if the offender received backlash, it might deter others from harassing them.

> *I think that me being confrontational about those who have harassed me deterred it [harassment]. It's kind of a public shaming type of thing. If you leave something that's harassing me, I will respond back like "does it make you feel good to be like this?" So I think it's nipping it in the bud [because] I don't really get it [harassment] anymore.* (I10)

Creators also mentioned attempting to make sure they do not accidentally disclose any private information or present location in their content. This includes measures such as posting pictures from events at a later date, keeping windows covered when filming so as not to expose their address, not using their real names online, and being vigilant about what personally identifiable information (PII) is available about them online.

However, despite their own efforts, participants also noted that it was difficult to keep their information from being exposed because they did not know they needed to take these precautions when they were starting out, and only felt the need to take action when their channel started to get more attention. Indeed, only one participant (I7) started their channel having taken precautions to keep themselves private. Moreover, participants mentioned the difficulty of keeping their information removed if it was once available. I04 recalls trying to remove their information from data broker sites, only for the information to reappear later.

> *I took a whole day trying to get myself bleached off of these [data broker] sites [data broker sites]. I think I was successful in doing that, but I don't know, because how a lot of these sites work ... maybe you're gone for a couple of weeks.*

Social media hygiene also applies to who the creator is associated with. As alluded to in Section 5.3.3, creators also felt that other creators could pose a threat in the event those creators have a falling out with the original creator, had bad intentions, or became embroiled in controversy. As such, creators also expressed needing to be cautious with whom they associated, as well as how close they became with other creators.

5.5 Participants' Requested Features

The features most participants requested were improvements to the platform's reporting and blocking functionality. Specifically, they wanted platforms to provide users with feedback regarding their reports (i.e., whether action was taken and reasons for this decision) and make blocking another user also block all other accounts associated with that account—both current and any future accounts they create—to make it more difficult to circumvent being blocked by creating new accounts or using an alternate account. Participants also suggested having an "Other" option when selecting the category a report falls under as well as the ability to attach a note when reporting to elaborate on why the content is being reported. Participants also spoke about wanting to be able to connect with a human when reaching out for support "rather than just filling out a form" (I07). A list of various suggestions participants stated they would like implemented can be found in Table A.2 in the appendix.

6 Discussion

6.1 Threats Arising from Intracommunity Dynamics

Prior work on online hate and harassment primarily addresses actions that reflect conventional conceptions of threats. In contrast, we found that participants' threat models included not only malicious actions but also *intracommunity dynamics*—the tensions, norms, and power dynamics present within the communities creators are situated in that

influence how creators interact (Walker and DeVito 2020). More specifically, participants sometimes viewed seemingly neutral or even positive actions from their own audience as a form of harassment. Participants cited growing and supporting their audience as a primary motivator for creating content; this was also a factor in how they decided what content to share, how to engage with their audience, and how to moderate their channels. Participants felt the responsibility to protect their audience from seeing potentially harmful comments, while still allowing their audience to exercise a degree of free expression. What constitutes an acceptable degree of moderation and what crosses the line into censorship seems to be determined by a combination of both the creator's own views and what they believed their audience expected. Most creators also wanted to build community with other creators, with some citing the importance of networking with higher-profile creators specifically as a way to access better support from the platform. At the same time, these same participants were also concerned that these connections could open them to further harassment by increasing their visibility. This highlights that part of risk management for creators is navigating multiple layers and types of intracommunity dynamics—both those within their own audience community and those within the broader community of creators.

6.1.1 Audience expectations

Many of the risks participants cited that are associated with these creator communities seem to stem from what audiences expect from creators. In addition to creating content that appeals to their intended audience, however, our participants also indicated a need to explicitly align themselves with the values of their audience. Failure to do so could result in backlash wherein audience members themselves may attempt to incite harassment against the creator in the form of "name-and-shame" campaigns. Online harassment is sometimes seen as a tool to enforce social norms and punish those who violate those norms (Blackwell et al. 2018; Take et al. 2024). Our findings indicate that for some creators, any association with brands or persons—including incidental or through personal consumption choices (e.g., products they are seen using, travel destinations)—may potentially be seen as a violation of norms. This is especially true if that brand or person is later seen as "problematic"; the association may then be viewed as a tacit endorsement of certain practices or values that audience members believe the brand or person represents. This potentially exposes the creator to future harassment—even if the creator was unaware of the implications of associating with or supporting the "problematic" brand or person.

Abuse stemming from backlash was not the only potential threat their community could pose. Participants cited as harassment interactions with fans that while seemingly positive or neutral, represented an overstepping of boundaries and an entitlement of the creator's time and attention. Indeed, though creators hold some celebrity status, the importance of being perceived as "accessible" (Valenzuela-García et al. 2023; Berryman and Kavka 2017)—especially as compared to traditional celebrities—can result

in audience members forming a parasocial relationship with creators wherein they also view the creator as a friend (Berryman and Kavka 2017). Though creators can attempt to manage this expectation, creating explicit boundaries between themselves and their audience can harm both their growth and their income potential. Moreover, as some audience members may view a creator's accessibility and relatability as a "requirement," if an established creator institutes *new* boundaries, it could itself be viewed by audience members as a transgression worthy of backlash.

While many of these risks may seem to originate in the interpersonal realm, their inextricable connection to creators' personal and financial circumstances constitutes a more substantive threat than for those who are not in this unique position. Creators occupy the roles of both idol and friend to their audience—that is, despite expecting creators to represent the average person, they are simultaneously put on a pedestal and expected to flawlessly represent their audiences' ideals. This dynamic between audience and creator, wherein creators are expected to be both above and equal to their audience, results in the community that creators seek to build and be a part of also becomes a potential source of harm.

These risks scale not only with the size of a creator and their own community, but also with the community's own reach. In communities where other creators are present, their audience may also pose a risk. Not only can creators mobilize their audience against other creators, but backlash can spill over to any creator known to associate with the "offending" creator. These risks are also difficult to predict, requiring understanding not only what values audiences expect, but also the actions and rhetoric that the audience may take as signaling support for opinions and practices that the audience supports and those they oppose. Moreover, these signals are not static, and because of the potential "long tail" of online content, older posts may be interpreted through a current lens. This was a phenomenon I10 experienced, receiving backlash for collaborating with a company prior to the company's controversy; thus, any present decision is risky in itself.

6.2 Persistent Systemic Risks

Our findings also broadly align with the findings of prior research on hate and harassment toward creators. Of note, our participants felt that reporting offending content had little impact nor remediation due to gaps in protection and lack of transparency from the platform (Gröber et al. 2024; Thomas et al. 2022; Reynolds and Hallinan 2024), and that moderation using platform-provided tools and social media hygiene are useful strategies in combating and mitigating hate; however, these practices were also seen as burdensome to creators and had a steep learning curve (Samermit et al. 2023; Soneji et al. 2024; Thomas et al. 2022). Moreover, some participants expressed concerns that moderation—both by creators and the platform—may be perceived as a form of censorship (Heung et al. 2024; Thomas et al. 2022).

Many of our findings around the gaps in protections afforded to those who create content that is shared online have been found in prior studies since at least 2017 (Blackwell et al. 2017), and are not unique to just "content creation" on social media platforms, but also on any public online platform that hosts content attributable to individuals or groups of individuals, including news platforms (Goyal, Park, and Vasserman 2022; Posetti et al. 2020) and collaborative projects like Wikipedia (Forte, Andalibi, and Greenstadt 2017). Existing literature on online abuse often situates itself in the HCI space, using findings to offer implications for designing online spaces or third-party tools meant to augment these online spaces (e.g., Scheuerman, Branham, and Hamidi 2018). Proposed solutions or improvements broadly advocate for third-party tooling and resources, or platform improvement based on feedback (Thomas et al. 2022; Uttarapong, Cai, and Wohn 2021). While third-party tooling has been useful for creators, changes to platforms' APIs may render them nonfunctional, as in the case of the tool Block Party (Perrigo 2023). While platforms have made improvements to tackle various types of toxic content on their platform (i.e., community notes to address misinformation), these improvements generally address only individual pieces of offending content, but not areas of structural breakdown that affords harassment (Lewis, Marwick, and Partin 2021) and creates difficulties in obtaining recourse when harassment occurs (Uttarapong, Cai, and Wohn 2021), nor do they mitigate specific creator concerns, e.g., more transparency about reports (Gröber et al. 2024).

The persistence of these challenges—similar solutions are repeatedly proposed and platforms repeatedly attempt to implement them—suggests the problem may require a different perspective. Creators' definitions of harassment are broader than those typically considered in prior literature (e.g., Valenzuela-García et al. 2023; Thomas et al. 2021), and the risk does not only stem from adversarial entities but also from those with competing interests (e.g., that of community members or the platform). This suggests that the current approach to tackling harassment may be too narrow and may overlook constraints and affordances that emerge from how subsystems interact within larger systems (Barile et al. 2012). For example, *systemic risks*, risks that arise from externalities created by interdependent systems, may contribute to the problem or impact the feasibility of existing solutions.

6.2.1 Lack of corporate incentives and legislative pressure

Content platforms—especially social media platforms—are for-profit entities, and as such, recommendations to platforms would need to consider how the suggested changes could impact profitability, especially in cases of publicly traded companies, which have fiduciary duties to shareholders (Zimmer 2021). In fact, this is what some of our participants contend is the reasoning behind preferential treatment toward very large creators. However, the profit motive may not be the only reason that platforms may not implement certain requested changes. Platforms also have to contend with potential legal liabilities around policy decisions, as they may be subject to the laws and regulations in the

different regions the platform can be accessed. This affects content moderation, because anything that may hold a platform accountable inherently opens the platform up to legal liabilities. For example, having community guidelines and terms of service that are more clear and specific as I09 suggested (see Table A.2, Appendix) may allow less room for moderator discretion in the platform's policies; however, it also may create a contractual commitment to be held to and scrutinized for—similar to, but distinct from, the conditions that necessitated Section 230[4] (Klonick 2018).

Effective legislation can be a strong motivator in incentivizing companies to adopt changes that would otherwise be unpopular for the company (Peeters 2020). However, technological innovation outpacing regulation is a well-documented problem (Margan 2025). The regulatory "pacing problem" stems not only from the inability to predict harms that will be caused by emerging technology but also policymakers' lack of domain knowledge, especially around specific online ecosystems. As a result, laws to regulate technology and the internet not only may not solve the problems the laws were intended to address, but also may create new problems.

The EU's Digital Services Act (DSA) and Online Safety Act (OSA) primarily address responsibilities of digital service providers regarding the trade and hosting of illicit goods and services on their platforms, and the duty of care platforms have toward underage users. Though the DSA does require that platforms provide transparency to users regarding the status of user reports and actions taken in response via clear and specific "statements of reason,"[5] based on transparency reports (European Commission, n.d.-b), outside of content deemed illegal, the information that platforms are required to provide to users is vague and minimal (European Commission, n.d.-a). While the DSA does encourage platforms to moderate to avoid penalties, Alkiviadou (2025) finds that in practice, much of this moderation is automated to comply with narrow time limits to remove illegal content imposed by law. In addition, without a clearly defined scope of what constitutes "illegal hate speech," these laws may be exploited by those who wish to stifle certain types of free expression (Alkiviadou 2025). Rather than adequately address the safety concerns of content creators, these laws also unintentionally exacerbate the censorship that marginalized groups already face and chill speech around controversial topics due to existing problems with platform moderation (see Section 2.4). This highlights how regulating allowed speech on platforms is ineffective in tackling hate and harassment.

Indeed, criticism has been levied on existing and proposed attempts to regulate social media platforms for similar ineffectiveness and negative impacts. For example, proposed age verification laws intended to keep children from accessing age-inappropriate material have been criticized not only for implications on privacy, freedom of speech, and

4. Section 230 of the Communications Decency Act grants platforms immunity from tortious liability for third-party user-generated content and moderating content in "good faith," but not liability due to breach of contract (Gregory 2021).

5. Article 17(1) and 17(4) of Regulation 2022/2065 (DSA).

their potential to allow for government surveillance (Lorenz 2026; Electronic Frontier Foundation et al. 2022), but also for counterintuitively requiring the collection of sensitive information from children (Weissman 2023; Macon 2026; Kelley and Schwartz 2023). Without existing trusted infrastructure to securely implement age verification, this puts all users at risk—including the children the law aimed to protect (Weissman and Flatley 2024). Recently, after Discord, an instant messaging and VOIP platform, implemented age verification in the UK and Australia to comply with the countries' newly introduced requirements, the third-party vendor used to handle customer service—including age-related appeals—suffered a data breach that resulted in a leak of photos, including government-issued identification users submitted during the appeals process (Alajaji and Baldwin 2026).

These examples illustrate a disconnect between legislative goals and technical realities due to lawmakers' limited understanding of online ecosystems and lack of technical expertise.

6.2.2 Creators face occupational hazards

Many of the risks associated with content creation parallel the risks of those faced by public figures and celebrities—indeed, as Valenzuela-García et al. (2023) note, "there is no clear separation between traditional celebrities and influencers." However, the protections creators receive are more similar to those who participate in the "gig economy," who are often not subject to the same protections and rights as workers in more traditional jobs and are subject to the design and policy choices of the platforms they use (Sannon, Sun, and Cosley 2022). In most jobs there are constraints to limit the safety risks workers must shoulder, and prior expectations in the tasks they will perform as well as potential risks they may face; however, the experiences of the creators in our study highlight that, unlike in other types of work, creators are unaware when they first start out of the tasks they will need to perform outside of creating content, as well as risk vectors they will have to navigate.

Navigating risk entails a high administrative burden for creators not only when facing harassment, but in preventing it to begin with. Our participants were often unaware of the preventive measures they needed to take before starting content creation, such as preemptively removing their PII off the internet or making sure sensitive information could not be inferred from their social media activities. This presents a safety issue, as personal information is already known to be hard to remove ordinarily (Take et al. 2022); however, when exposed to malicious actors, it may be impossible to permanently remove—or worse, attempts to request removal could expose additional information as well as revive or escalate harassment (Take et al. 2024).

Despite the importance of creators to the platform—social media is co-created by creators, and a platform's value is reliant on creators making content on the platform (Deloitte 2025)—much of the burden of ensuring a safe work environment is put on the

creator. The measures the platform has in place are insufficient to ease the administrative burden creators face when attempting to navigate the risks of content creation. Although the platform does provide tools that participants found useful, participants felt that they lacked guidance on how best to utilize them. Moreover, the platform-provided tools are remedial—that is, designed for creators to respond to incidents after they have occurred—and onerous at scale. While platform support exists, our participants' responses indicate that the process of obtaining support from the platform entails a greater administrative burden due to friction in the reporting process and difficulties for small creators in contacting a support agent.

Content can unexpectedly "go viral," exposing the creator to not only a massive increase in visibility, but also risk of harassment (Samermit et al. 2023). In fact, our participants often did not start content creation expecting exponential growth. The unpredictability of risks associated with content creation and the low barrier to entry presents a safety risk not only to those who identify as content creators but also to users who publicly share "content" on social media platforms. For those users who are not currently generating an income through their content, this can be especially devastating: there have been documented cases of these users losing employment after a post they had made reached beyond their intended audience—such as those who had been targeted by "name-and-shame" campaigns. This indicates a need for platforms to provide support to potential creators from the point of sign-up, rather than only after they have begun to amass a following—and the increased risks that come with it.

6.3 Recommendations

Creators, platforms, and even researchers are situated within larger systems and cannot be fully divorced from these systems due to the ways entities within a system interact with and influence each other. As such, though individual groups can make improvements within their own subsystems, solutions for the broader problem at hand require collaboration between relevant subsystems. In the following, we make suggestions for ways individual stakeholders can take action in working toward improvements.

6.3.1 Platforms

Creators are a fundamental part of the supply chain that generates profit for platforms. The value of platforms is dependent on creators continuing to provide original content to serve users. Some creators choose to no longer create content due to harassment, which can be seen as a loss of a value-generating resource. As such, we argue that setting prospective creators up for success by providing them support from initial onboarding and throughout their tenure as a content creator on the platform serves the interests of the platform.

While we recognize that platforms do not want to "chill" content that is experiencing sudden growth, for creators, sudden "virality" is an inflection point where there is a spike

in difficulty managing engagement due to volume alone and an exponential increase in abusive content targeted at them. Platforms could do more to prepare creators for unexpected growth and mitigate potential harm resulting from this growth. This could take the form of nudges that caution and recommendations that highlight and encourage best practices for personal safety and security at relevant growth milestones (Moynihan, Herd, and Harvey 2015); for example, recommending that new users have separate personal and professional accounts with separate associated emails during the sign-up process, or reminding users not to share highly identifiable information upon first upload. Though these suggestions do not address problems around sudden, unexpected growth, they can help creators start off with a good foundation of knowledge to avoid common mistakes with long-term consequences—and are also resources that participants stated they would have liked to have had when starting out.

In addition, platforms could consider features that other platforms claim to have and that users found would be useful, such as the highly requested "universal" block, which Meta currently claims exists on Threads and Instagram (Instagram, n.d.), two Meta platforms. The universal block would block not only a user, but all other current and future accounts associated with that user. Unfortunately, independent auditing has found that the universal block feature Meta claimed to implement for Instagram appears to be defective or no longer in place (Béjar et al. 2025). However, this type of universal block feature could be implemented more robustly in the future.

6.3.2 Creators

Creators have consistently voiced their desire to be in community with other creators, both in this study and across research on creators more broadly (Scheuerman, Branham, and Hamidi 2018; Samermit et al. 2023; Sannon et al. 2023; Thomas et al. 2022). Indeed, the value of peer support as emotional support as well as mentorship has been documented in a variety of settings (Van Maanen and Schein 1979; Fineman 2014; L. Davidson et al. 2006). Though creators may have unique challenges due to not only the types of content they create, but the intersection of their identity (Heung et al. 2024; Scheuerman, Branham, and Hamidi 2018), creators still do share common lived experiences and understanding that make other creators better positioned to provide support than non-creators.

This support could not only be emotional support or mentorship, but also take the form of collective action and collective representation (Reynolds and Hallinan 2024). Forming even a loosely connected group with similar shared interests is beneficial to coordinating various forms of advocacy, such as collective bargaining and pushes for policy change (Kochan et al. 2023; Lin 2022; Lamannis 2023). Creators should look to examples of how others in the gig economy have organized for better conditions, e.g., app-based delivery drivers (Meditz 2025) and rideshare drivers (Dubal 2019).

6.3.3 Researchers

Although researchers cannot address all the constraints highlighted in Section 6.2, which have kept things static over the past decade, researchers are able to provide evidence and guidance that policymakers need to create better protections. We suggest researchers reach out to policymakers, especially at local and regional levels, to understand the types of research evidence that are missing or most helpful for developing sound policy on related issues.

We recommend that future study designs account for variation in how participants may interpret or conceptualize terms and concepts, to avoid overlooking dimensions that fall outside traditional definitions. One way this can be achieved is through pilot testing to help identify potential priming effects (Ruel, Wagner, and Gillespie 2016). Alternatively, researchers could include a question asking participants to define key concepts (Charmaz 2014). Where the goal of research is to develop actionable suggestions, we suggest researchers consider a participatory approach, where researchers work with participants not as subjects but as co-researchers in their own right. Not only does this allow research to be shaped by expertise grounded in lived experience, but because knowledge is shared bidirectionally, this also gives those in the community tools to advocate for themselves (Mergler 1987; Banks et al. 2018; Glass and Newman 2015).

6.3.4 Legislators

Although creators are essential to platform profitability, platforms fail to provide adequate protection to their creators. We recommend that lawmakers work with creators toward enhancing labor protections for creators similar to those that protect those working in the gig economy. As creators are also end-users, this would also help to protect all users. Such protections should, at a starting point, ensure that creators have safe working conditions, such as being provided resources and recourse when facing harassment.

In addition, we also suggest policymakers take a user-centered design approach to understand the intricacies of any socio-technical system to inform legislation. This includes not only involving individuals with technical expertise, expertise in relevant socio-technical ecosystems, and end users, but also existing research on the needs of users of these systems.

6.3.5 Community

We hesitate to make any specific recommendations on how community members should conduct themselves, as our participants indicated that each community and creator has their own preferences for how their community behaves and engages with them and their content. However, the tension between creators and audiences' expectations for them holds significant weight within the threat model of creators. The problem of parasociality predates the existence of content creators. Unlike other public media figures

and traditional celebrities who can maintain distance from their fans, however, engaging with one's audience is a necessary part of the job, and community building is a motivation for being a content creator. These are not necessarily people you want to ban.

How to best support creators in managing these intracommunity risks is an area for further research. One research direction we suggest is investigating how platform design can both incentivize and discourage the audience interactions that creators find harmful. Some platforms, such as Twitch (Twitch, n.d.) or Reddit (Reddit, n.d.), allow creators and communities to set their own clearly defined and visible rules and expectations of conduct for their communities. However, the ability to set boundaries does not eliminate the risk creators face from their community and may actually realize the risk.

While we cannot offer suggestions for the root problem, we propose some suggestions that may address immediate creator desires around community. The creators we spoke to wanted to connect with other creators for emotional support due to a shared understanding of the problems creators face, but did not know who they could trust. Platforms could consider a model similar to that used by Reddit's volunteer moderators. Similarly, many multiplayer video games have programs that allow players to opt-in to volunteer as player moderators. For example, Dota 2's and Counter-Strike's Overwatch systems (Dota 2 2021), RuneScape's Player Moderator program (RuneScape, n.d.), or Threads' and Twitter/X's crowdsourcing of community notes (Threads, n.d.; Twitter, n.d.). Platforms or even a third party could introduce a program where creators could opt-in, and after determined trustworthy, anonymously provide emotional support to anonymous creators who need it.

In addition, the burden is still on the creator to manage the inflection point where creators shift from wanting to be accessible to grow and when their channel scales enough to need to enforce boundaries, and also be able to know what boundaries are needed. One potential way to address this problem could be to publicly introduce a feature that "activates" certain settings by default at growth milestones or, if brigading is detected, allows creators time to adjust, as well as avoid scrutiny because the settings were implemented by the platform.

References

Alajaji, Rindala, and Samantha Baldwin. 2026. "Discord Voluntarily Pushes Mandatory Age Verification Despite Recent Data Breach." Electronic Frontier Foundation, February 12, 2026. https://www.eff.org/deeplinks/2026/02/discord-voluntarily-pushes-mandatory-age-verification-despite-recent-data-breach.

Alkiviadou, Natalie. 2025. "Platform Liability, Hate Speech and the Fundamental Right to Free Speech." *Information & Communications Technology Law* 34 (2): 207–17. https://doi.org/10.1080/13600834.2024.2411799.

Banks, Sarah, Andrea Armstrong, Anne Bonner, Yvonne Hall, Patrick Harman, Luke Johnston, Clare Levi, Kath Smith, and Ruth Taylor. 2018. "Between Research and Community Development: Negotiating a Contested Space for Collaboration and Creativity." In *Co-Producing Research: A Community Development Approach,* edited by Angie Hart, Kate Pahl, Paul Ward, and Sarah Banks, 21–48. Bristol University Press. https://doi.org/10.46692/9781447340775.004.

Barile, Sergio, Jaqueline Pels, Francesco Polese, and Marialuisa Saviano. 2012. "An Introduction to the Viable Systems Approach and Its Contribution to Marketing." *Journal of Business Market Management* 5 (2): 54–78. https://hdl.handle.net/10419/66007.

Béjar, Arturo, Molly Rose Foundation, Cybersecurity for Democracy, Parents for Safe Online Spaces, Fairplay, and Heat Initiative. 2025. "Teen Accounts, Broken Promises: How Instagram is Failing to Protect Minors," September 23, 2025. https://fairplayforkids.org/wp-content/uploads/2025/09/Teen-Accounts-Broken-Promises-How-Instagram-is-failing-to-protect-minors.pdf.

Berryman, Rachel, and Misha Kavka. 2017. "'I Guess A Lot of People See Me as a Big Sister or a Friend': The Role of Intimacy in the Celebrification of Beauty Vloggers." *Journal of Gender Studies* 26 (3): 307–20. https://doi.org/10.1080/09589236.2017.1288611.

Bhuiyan, Shehreen Amin, and Hasan Muhammad Baniamin. 2025. "Intent vs Impact: Ramifications of Digitalization on the Administrative Burden." *Public Policy and Administration,* https://doi.org/10.1177/09520767251348541.

Blackwell, Lindsay, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. 2018. "When Online Harassment Is Perceived as Justified." *Proceedings of the International AAAI Conference on Web and Social Media* 12 (1). https://doi.org/10.1609/icwsm.v12i1.15036.

Blackwell, Lindsay, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. "Classification and Its Consequences for Online Harassment: Design Insights from HeartMob." *Proceedings of the ACM on Human-Computer Interaction* 1:1–19. https://doi.org/10.1145/3134659.

Chadha, Kalyani, Linda Steiner, Jessica Vitak, and Zahra Ashktorab. 2020. "Women's Responses to Online Harassment." *International Journal of Communication* 14:239–57. https://ijoc.org/index.php/ijoc/article/view/11683.

Charmaz, Kathy. 2014. *Constructing Grounded Theory.* SAGE Publications.

Dastagir, Alia E. 2021. "'The Internet Is Not a Game. ... This Stuff Really Hurts.' Respected Developer Who Was Bullied Online Dies by Suicide." USA TODAY, July 23, 2021. https://www.usatoday.com/story/tech/2021/07/23/how-toxic-online-cultures-trolling-and-bullying-contribute-suicide/8042846002/.

Davidson, Larry, Matthew Chinman, David Sells, and Michael Rowe. 2006. "Peer Support Among Adults with Serious Mental Illness: A Report from the Field." *Schizophrenia Bulletin* 32 (3): 443–50. https://doi.org/10.1093/schbul/sbj043.

Davidson, Thomas, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. "Automated Hate Speech Detection and the Problem of Offensive Language." *arXiv,* https://doi.org/10.48550/arXiv.1703.04009.

Deloitte. 2025. "Creator Lifetime Value in a Content Creator Ecosystem," March 25, 2025. https://www.deloitte.com/us/en/services/consulting/articles/content-creator-ecosystem-and-influencer-loyalty.html.

Dota 2. 2021. "Overwatch Arrives in Dota 2021," January 27, 2021. https://www.dota2.com/newsentry/3025824821114909461.

Duarte, Natasha, Emma Llanso, and Anna Loup. 2018. "Mixed Messages? The Limits of Automated Social Media Content Analysis." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency,* edited by Sorelle A. Friedler and Christo Wilson, vol. 81. https://proceedings.mlr.press/v81/duarte18a.html.

Dubal, Veena. 2019. "Why the Uber Strike Was a Triumph." Slate, May 10, 2019. https://slate.com/technology/2019/05/uber-strike-victory-drivers-network.html.

Electronic Frontier Foundation et al. 2022. "Re: Opposition to S. 3663's Threats to Minors' Privacy and Safety Online," November 28, 2022. https://www.eff.org/node/107521.

European Commission. n.d.-a. "Additional Explanation For Statement Attributes." DSA Transparency Database. https://transparency.dsa.ec.europa.eu/page/additional-explanation-for-statement-attributes#information-on-the-facts-and-circumstances-relied-on-in-taking-the-decision.

———. n.d.-b. "DSA Transparency Database." DSA Transparency Database. https://transparency.dsa.ec.europa.eu/.

Fineman, Martha Albertson. 2014. "Vulnerability, Resilience, and LGBT Youth." *Temple Political & Civil Rights Law Review* 23 (2): 307–30.

Forte, Andrea, Nazanin Andalibi, and Rachel Greenstadt. 2017. "Privacy, Anonymity, and Perceived Risk in Open Collaboration: A Study of Tor Users and Wikipedians." In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing,* 1800–1811. https://doi.org/10.1145/2998181.2998273.

Glass, Ronald David, and Anne Newman. 2015. "Ethical and Epistemic Dilemmas in Knowledge Production: Addressing Their Intersection in Collaborative, Community-Based Research." *Theory and Research in Education* 13 (1). https://doi.org/10.1177/1477878515571178.

Google. n.d. "Submit a Copyright Counter Notification." Google Support. https://support.google.com/youtube/answer/2807684.

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." *Big Data & Society* 7 (1). https://doi.org/10.1177/2053951719897945.

Goyal, Nitesh, Leslie Park, and Lucy Vasserman. 2022. "'You Have to Prove the Threat is Real': Understanding the Needs of Female Journalists and Activists to Document and Report Online Harassment." In *CHI Conference on Human Factors in Computing Systems,* 1–17. https://doi.org/10.1145/3491102.3517517.

Grassegger, Julia, and Hannes Angwin. 2017. "Facebook's Secret Censorship Rules Protect White Men from Hate Speech but Not Black Children." ProPublica, June 28, 2017. https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms.

Gregory, Vellos. 2021. "Is It Possible to Circumvent § 230 with Contract Claims?" Georgetown Journal of Law & Public Policy Legal Blog, December 8, 2021. https://www.law.georgetown.edu/public-policy-journal/blog/is-it-possible-to-circumvent-%c2%a7-230-with-contract-claims/.

Gröber, Lea, Waleed Arshad, Shanza, Angelica Goetzen, Elissa M. Redmiles, Maryam Mustafa, and Katharina Krombholz. 2024. "'I Chose to Fight, Be Brave, and to Deal with It': Threat Experiences and Security Practices of Pakistani Content Creators." In *33rd USENIX Security Symposium (USENIX Security 24),* 19–36. https://www.usenix.org/conference/usenixsecurity24/presentation/gr%C3%B6ber-content-creators.

Heung, Sharon, Lucy Jiang, Shiri Azenkot, and Aditya Vashistha. 2024. "'Vulnerable, Victimized, and Objectified': Understanding Ableist Hate and Harassment Experienced by Disabled Content Creators on Social Media." In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3613904.3641949.

Holton, Avery E., Valérie Bélair-Gagnon, Diana Bossio, and Logan Molyneux. 2023. "'Not Their Fault, but Their Problem': Organizational Responses to the Online Harassment of Journalists." *Journalism Practice* 17 (4): 859–74. https://doi.org/10.1080/17512786.2021.1946417.

Instagram. n.d. “What Happens When You Block Someone on Instagram.” Instagram Help Center. https://help.instagram.com/447613741984126.

Jhaver, Shagun, Larry Chan, and Amy Bruckman. 2018. “The View from the Other Side: The Border Between Controversial Speech and Harassment on Kotaku in Action.” *First Monday* 23 (2). https://doi.org/10.5210/fm.v23i2.8232.

Jiang, Jialun Aaron, Peipei Nie, Jed R. Brubaker, and Casey Fiesler. 2023. “A Trade-off-Centered Framework of Content Moderation.” *ACM Transactions on Computer-Human Interaction* 30 (1): 1–34. https://doi.org/10.1145/3534929.

Keighley, Rachel. 2022. “Hate Hurts: Exploring the Impact of Online Hate on LGBTQ+ Young People.” *Women & Criminal Justice* 32 (1–2): 29–48. https://doi.org/10.1080/08974454.2021.1988034.

Kelley, Jason, and Adam Schwartz. 2023. “Age Verification Mandates Would Undermine Anonymity Online.” Electronic Frontier Foundation, March 10, 2023. https://www.eff.org/deeplinks/2023/03/age-verification-mandates-would-undermine-anonymity-online.

Klonick, Kate. 2018. “The New Governors: The People, Rules, and Processes Governing Online Speech.” *Harvard Law Review* 131 (6). https://harvardlawreview.org/print/vol-131/the-new-governors-the-people-rules-and-processes-governing-online-speech/.

Kochan, Thomas A., Janice R. Fine, Kate Bronfenbrenner, Suresh Naidu, Jacob Barnes, Yaminette Diaz-Linhart, John Kallas, et al. 2023. “An Overview of US Workers’ Current Organizing Efforts and Collective Actions.” *Work and Occupations* 50 (3): 335–50. https://doi.org/10.1177/07308884231168793.

Lamannis, Mariagrazia. 2023. “Collective Bargaining in the Platform Economy: A Mapping Exercise of Existing Initiatives.” *Social Science Research Network,* no. 4404691, https://doi.org/10.2139/ssrn.4404691.

Lewis, Rebecca, Alice E. Marwick, and William Clyde Partin. 2021. “‘We Dissect Stupidity and Respond to It’: Response Videos and Networked Harassment on YouTube.” *American Behavioral Scientist* 65 (5): 735–56. https://doi.org/10.1177/0002764221989781.

Lin, Lefeng. 2022. “Power Resources and Workplace Collective Bargaining: Evidence from China.” *Journal of Chinese Sociology* 9 (1). https://doi.org/10.1186/s40711-022-00178-x.

Lorenz, Taylor. 2026. “Congress Is Considering Abolishing Your Right to Be Anonymous Online.” The Intercept, March 5, 2026. https://theintercept.com/2026/03/05/kosa-online-age-verification-free-speech-privacy/.

Macon, Ken. 2026. "Arizona Bill Would Require ID Checks to Use a Weather App." Reclaim the Net, February 16, 2026. https://reclaimthenet.org/arizona-bill-would-require-id-checks-to-use-a-weather-app.

Malinsky, Gili. 2024. "57% of Gen Zers Want to be Influencers—But 'It's Constant, Monday through Sunday,' Says Creator." CNBC, September 14, 2024. https://www.cnbc.com/2024/09/14/more-than-half-of-gen-z-want-to-be-influencers-but-its-constant.html.

Margan, Srivathsan Karanai. 2025. "The Pacing Problem Unplugged Part 1." American Academy of Actuaries, January 1, 2025. https://actuary.org/article/the-pacing-problem-unplugged-part-1/.

Mariconti, Enrico, 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." *Proceedings of the ACM on Human-Computer Interaction* 3:1–21. https://doi.org/10.1145/3359309.

Meditz, Stephanie G. 2025. "Council OKs Delivery Worker Protections." Queens Chronicle, July 17, 2025. https://www.qchron.com/editions/queenswide/council-oks-delivery-worker-protections/article_859ba579-338a-52e7-b3d5-70f0abc24325.html.

Mergler, Donna. 1987. "Worker Participation in Occupational Health Research: Theory and Practice." *International Journal of Health Services: Planning, Administration, Evaluation* 17 (1): 151–67. https://doi.org/10.2190/FPWF-C2ET-Q4DB-NMNQ.

Moynihan, Donald, Eric Giannella, Pamela Herd, and Julie Sutherland. 2022. "Matching to Categories: Learning and Compliance Costs in Administrative Processes." *Journal of Public Administration Research and Theory* 32 (4): 750–64. https://doi.org/10.1093/jopart/muac002.

Moynihan, Donald, Pamela Herd, and Hope Harvey. 2015. "Administrative Burden: Learning, Psychological, and Compliance Costs in Citizen-State Interactions." *Journal of Public Administration Research and Theory* 25 (1): 43–69. https://doi.org/10.1093/jopart/muu009.

Nadim, Marjan, and Audun Fladmoe. 2021. "Silencing Women? Gender and Online Harassment." *Social Science Computer Review* 39 (2): 245–58. https://doi.org/10.1177/0894439319865518.

Nadler, Anthony, Matthew Crain, and Joan Donovan. 2018. *Weaponizing the Digital Influence Machine.* Data & Society. https://datasociety.net/wp-content/uploads/2018/10/DS_Digital_Influence_Machine.pdf.

Pastor, Ada Romero. 2024. "The Censorship of LGBTQ+ Content Online Corresponds with Declines in Freedom for Everyone." Tech Policy Press, October 30, 2024. https://www.techpolicy.press/the-censorship-of-lgbtq-content-online-corresponds-with-declines-in-freedom-for-everyone/.

Peeters, Rik. 2020. "The Political Economy of Administrative Burdens: A Theoretical Framework for Analyzing the Organizational Origins of Administrative Burdens." *Administration & Society* 52 (4): 566–92. https://doi.org/10.1177/0095399719854367.

Perrigo, Billy. 2023. "Her App Blocked Harassment On Twitter. Elon Musk Killed It." Time, June 2, 2023. https://time.com/6284494/block-party-twitter-tracy-chou-elon-musk/.

Posetti, Julie, Nermine Aboulez, Kalina Bontcheva, Jackie Harrison, and Silvio Waisbord. 2020. *Online Violence Against Women Journalists: A Global Snapshot of Incidence and Impacts.* Paris, France: UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000375136.

Reddit. n.d. "Reddit Rules." https://redditinc.com/policies/reddit-rules.

Reynolds, CJ, and Blake Hallinan. 2024. "User-Generated Accountability: Public Participation in Algorithmic Governance on YouTube." *New Media & Society* 26 (9): 5107–29. https://doi.org/10.1177/14614448241251791.

Ruel, Erin, William Edward Wagner, and Brian Joseph Gillespie. 2016. "Pretesting and Pilot Testing." In *The Practice of Survey Research: Theory and Applications,* 101–19. SAGE Publications. https://doi.org/10.4135/9781483391700.n6.

RuneScape. n.d. "Moderators and Community Helpers." Jagex. https://support.runescape.com/hc/en-gb/articles/209270969-Moderators-and-Community-helpers.

Samermit, Patrawat, Anna Turner, Patrick Gage Kelley, Tara Matthews, Vanessia Wu, Sunny Consolvo, and Kurt Thomas. 2023. "'Millions of People Are Watching You': Understanding the Digital-Safety Needs and Practices of Creators." In *32nd USENIX Security Symposium (USENIX Security 23),* 5629–45. https://www.usenix.org/conference/usenixsecurity23/presentation/samermit.

Sannon, Shruti, Billie Sun, and Dan Cosley. 2022. "Privacy, Surveillance, and Power in the Gig Economy." In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems,* 1–15. https://doi.org/10.1145/3491102.3502083.

Sannon, Shruti, Jordyn Young, Erica Shusas, and Andrea Forte. 2023. "Disability Activism on Social Media: Sociotechnical Challenges in the Pursuit of Visibility." In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3544548.3581333.

Scheuerman, Morgan Klaus, 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* 2:1–27. https://doi.org/10.1145/3274424.

Soneji, Ananta, Vaughn Hamilton, Adam Doupé, Allison McDonald, and Elissa M. Redmiles. 2024. "'I Feel Physically Safe but Not Politically Safe': Understanding the Digital Threats and Safety Practices of OnlyFans Creators." In *33rd USENIX Security Symposium (USENIX Security 24),* 1–18. https://www.usenix.org/conference/usenixsecurity24/presentation/soneji.

Take, Kejsi, Kevin Gallagher, Andrea Forte, Damon McCoy, and Rachel Greenstadt. 2022. "'It Feels Like Whack-a-Mole': User Experiences of Data Removal from People Search Websites." *Proceedings on Privacy Enhancing Technologies* 2022 (3): 159–78. https://doi.org/10.56553/popets-2022-0067.

Take, Kejsi, Victoria Zhong, Chris Geeng, Emmi Bevensee, Damon McCoy, and Rachel Greenstadt. 2024. "Stoking the Flames: Understanding Escalation in an Online Harassment Community." *Proceedings of the ACM on Human-Computer Interaction* 8:1–23. https://doi.org/10.1145/3641015.

Thomas, Kurt, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, et al. 2021. "SoK: Hate, Harassment, and the Changing Landscape of Online Abuse." In *2021 IEEE Symposium on Security and Privacy (SP),* 247–67. https://doi.org/10.1109/SP40001.2021.00028.

Thomas, Kurt, 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 *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* https://doi.org/10.1145/3491102.3501879.

Threads. n.d. "Introducing Community Notes - Adding Context to Posts." Meta. https://www.meta.com/technologies/community-notes/.

Twitch. n.d. "Setting Up Moderation for Your Twitch Channel." https://help.twitch.tv/s/article/setting-up-moderation-for-your-twitch-channel.

Twitter. n.d. "About Community Notes on X." X Help. https://help.x.com/en/using-x/community-notes.

U.S. Attorney's Office, District of Kansas. 2019. "Ohio Gamer Pleads Guilty in Swatting That Caused a Death." U.S. Department of Justice, April 3, 2019. https://www.justice.gov/usao-ks/pr/ohio-gamer-pleads-guilty-swatting-caused-death.

Uttarapong, Jirassaya, Jie Cai, and Donghee Yvette Wohn. 2021. "Harassment Experiences of Women and LGBTQ Live Streamers and How They Handled Negativity." In *Proceedings of the 2021 ACM International Conference on Interactive Media Experiences,* 7–19. https://doi.org/10.1145/3452918.3458794.

Valenzuela-García, Noelia, Diego J. Maldonado-Guzmán, Andrea García-Pérez, and Cristina Del-Real. 2023. "Too Lucky to Be a Victim? An Exploratory Study of Online Harassment and Hate Messages Faced by Social Media Influencers." *European Journal on Criminal Policy and Research* 29 (3): 397–421. https://doi.org/10.1007/s10610-023-09542-0.

Van Maanen, John, and Edgar H. Schein. 1979. "Toward a Theory of Organizational Socialization." *Research in Organizational Behavior* 1:209–64. https://api.semanticscholar.org/CorpusID:11868398.

Veale, Kevin. 2020. "Introduction: The Breadth of Harassment Culture and Contextualising Gamergate." In *Gaming the Dynamics of Online Harassment,* edited by Kevin Veale, 1–33. Springer International Publishing. https://doi.org/10.1007/978-3-030-60410-3_1.

Walker, Ashley Marie, and Michael Ann DeVito. 2020. ""'More Gay' Fits in Better": Intracommunity Power Dynamics and Harms in Online LGBTQ+ Spaces." In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems,* 1–15. https://doi.org/10.1145/3313831.3376497.

Weissman, Shoshana. 2023. "Age-Verification Legislation Discourages Data Minimization, Even When Legislators Don't Intend That." R Street, May 24, 2023. https://www.rstreet.org/commentary/age-verification-legislation-discourages-data-minimization-even-when-legislators-dont-intend-that/.

Weissman, Shoshana, and Maureen Flatley. 2024. "25 Percent of Kids Will Face Identity Theft Before Turning 18. Age-Verification Laws Will Make This Worse." R Street, July 25, 2024. https://www.rstreet.org/commentary/25-percent-of-kids-will-face-identity-theft-before-turning-18-age-verification-laws-will-make-this-worse/.

Wohn, Donghee Yvette, and Guo Freeman. 2020. "Audience Management Practices of Live Streamers on Twitch." In *Proceedings of the 2020 ACM International Conference on Interactive Media Experiences,* 106–16. https://doi.org/10.1145/3391614.3393653.

Zimmer, Scott. 2021. "Stakeholder Theory and Analysis." EBSCO, April 1, 2021. https://www.ebsco.com/research-starters/business-and-management/stakeholder-theory-and-analysis.

Authors

Victoria Zhong is a PhD Candidate at New York University. Email: vzhong@nyu.edu.

Meghna Nair is a PhD Student at New York University

Susan McGregor is a Professor at Columbia University

Damon McCoy is an Associate Professor at New York University

Rachel Greenstadt is an Associate Professor at New York University

Acknowledgments

We would like to thank the creators who participated in this study, as well as Ashley Marie Walker and the anonymous reviewers who provided constructive feedback.

Data availability statement

Not applicable.

Funding statement

The study was partially supported by The Google Cyber NYC Institutional Research Program. The first author was supported by a US Department of Education GAANN Fellowship (#P200A210096).

Ethical standards

Our Institutional Review Board (IRB) has reviewed and approved our study protocol, interview materials, consent, and recruitment procedures. The topic of the study was clearly stated in the recruitment study. Prior to the interview, we provided all participants with a consent notice via email, and during the interview, we asked participants to reconfirm their consent to participate in the study, as well as for us to record the interviews and retain their emails for further contact. Participants were allowed to decline to answer any question if they were not comfortable answering, and it would not affect their eligibility to receive the compensation. Interview participants were compensated with a $20 Amazon gift card after completing the interview.

Keywords

YouTube content creators; hate and harassment; online communities; content moderation tools; platform governance; intra-community dynamics; interviews.

Appendices

Appendix A: Tables

Table A.1: **Demographics and channel status of participants.**

IID	Content Type	Pronouns	Status*	Location
I01	BookTuber, Commentary	She/Her	Silver	US
I02	Horror, Video Games, Let's Play	He/Him	Silver	US
I03	Commentary	She/Her	None	CA
I04	Anime, Video Essays, Documentary	They/Them	Silver	US
I05	Fitness	He/Him	None	US
I06	Science Documentary	He/Him	Silver	CA
I07	Sitcom Retrospectives, Documentary	He/Him	Silver	CA
I08	Science Educator	She/Her	None	US
I09	Video Games, Video Essay	Any/All	Silver	US
I10	Crafting	She/Her	None	US
I11	Commentary	No Response	None	US
I12	BookTuber	He/Him	Silver	US
I13	Commentary	He/Him	Gold	US
I14	Internet History, Documentary	He/Him	None	US
I15	Crafting, Commentary	They/Them	Silver	US
I16	Science Educator	None	Silver	US
I17	Film, Video Essay	She/Her	Silver	US
I18	Literature	He/Him	None	US
I19	Philosophy, Religion	He/Him	Silver	US

* YouTube subscriber status via the minimum number of subscribers. Silver: >100K; Gold: >1M.

Table A.2: **Participants' desired features and resources.**

Problem Area	Suggestions
Reporting	1. Recieving feedback on reports made, including if action was(n't) taken (I04, 07, 10, 11, 15, 16), and for those who have consistently filed valid reports (I09) 2. More platform transparency about their decisions regarding reports (I13) and platform moderation (I09) 3. Option to appeal report (I14) 4. Ability to select an "Other" category as well as multiple categories (I11) 5. Ability to attach a note or comment to report (I10, 11, 15) 6. Notifying the person who was reported so they can be aware of their behavior (I14) 7. More streamlined reporting process (I04) 8. Ability to mass report content (I07)
Blocking/Hiding Users	1. Blocking/hiding a user from the channel affects all other current and future accounts of that user as well (I02, 04, 05, 06, 07, 14, 15) 2. Ability to utilize "blocklists" to block a list of users other creators deem worthy of blocking (I04, 06) 3. Blocking a user also restricts that user's ability to see the person who blocked their content (I05)
Platform Support/Policies	1. Easier access to a human when reaching out to the platform for support (I03, 12, 13) 2. Clear and specific community guidelines on what is or is not allowed (I09) 3. Dedicated staff that primarily handles harassment and community guideline related problems (I19) and provides personal support for those dealing with harassment (I16) 4. Not allowing banned users to circumvent bans by creating new accounts, banning new accounts the user attempts to create, or shadowbanning them (I07, 9) 5. Restricting the ability of frequent offenders to comment (I09, 19) 6. Limiting the level of interaction a user can have with a creator (I01), i.e., the number of times a user can comment on a specific person's content (I03)
Resources	1. Access to resources on being comfortable with having an online persona (I10) 2. Access to resources on strategies for handling and "weathering" harassment (I02, 07) 3. Access to mental health resources/services (I04, 07)
Misc.	1. Ability to choose the audience to whom the creator's content gets recommended (I05)