Hate Raids on Twitch: Understanding Real-Time Human-Bot Coordinated Attacks in Live Streaming Communities

JIE CAI, Pennsylvania State University, USA
SAGNIK CHOWDHURY, New Jersey Institute of Technology, USA
HONGYANG ZHOU, New York University, USA
DONGHEE YVETTE WOHN, New Jersey Institute of Technology, USA

Online harassment and content moderation have been well-documented in online communities. However, new contexts and systems always bring new ways of harassment and need new moderation mechanisms. This study focuses on *hate raids*, a form of group attack in real-time in live streaming communities. Through a qualitative analysis of hate raids discussion in the Twitch subreddit (r/Twitch), we found that (1) hate raids as a human-bot coordinated group attack leverages the live stream system to attack marginalized streamers and other potential groups with(out) breaking the rules, (2) marginalized streamers suffer compound harms with insufficient support from the platform, (3) moderation strategies are overwhelmingly technical, but streamers still struggle to balance moderation and participation considering their marginalization status and needs. We use affordances as a lens to explain how hate raids happens in live streaming systems and propose *moderation-by-design* as a lens when developing new features or systems to mitigate the potential abuse of such designs.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**; *Empirical studies in HCI*.

Additional Key Words and Phrases: content moderation; platform governance; live streaming; marginalized group; group attack; human-bot collaboration; harassment; affordances

ACM Reference Format:
Jie Cai, Sagnik Chowdhury, Hongyang Zhou, and Donghee Yvette Wohn. 2023. Hate Raids on Twitch: Understanding Real-Time Human-Bot Coordinated Attacks in Live Streaming Communities. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 342 (October 2023), 28 pages. https://doi.org/10.1145/3610191

1 INTRODUCTION

Online abuse, also termed online harassment, is defined as "pervasive or severe targeting of an individual or group online through harmful behavior" [1]. It is a prevalent and persistent problem for many online communities, from online forums decades ago to social media platforms like Facebook, Instagram, Reddit, and Twitter, to recent novel communities with real-time interaction like Twitch, Clubhouse, and Metaverse. It entails multiple harms to users [88] and those who deal with content [23, 96]. Victims of harassment consider harassment an ongoing event and need various forms of support [34].

In 2021, Twitch, a leading live streaming platform that provides multimodal interaction between broadcasters (streamers) and the audience (viewers), experienced a boycott by its users, primarily

Authors' addresses: Jie Cai, Pennsylvania State University, University Park, USA, jie.cai@psu.edu; Sagnik Chowdhury, New Jersey Institute of Technology, Newark, USA, sc25@njit.edu; Hongyang Zhou, New York University, New York, USA, hz2187@nyu.edu; Donghee Yvette Wohn, New Jersey Institute of Technology, Newark, USA, yvettewohn@gmail.com.

2573-0142/2023/10-ART342 $15.00
https://doi.org/10.1145/3610191

marginalized streamers, complaining about its inability to handle hate raids on its platform. Hate raids is a unique term developed for Twitch communities, originating from the "Raids" feature, which allows a streamer, after their stream, to send their viewers to other streamers' chatrooms to support each other's community. Hate raids occur when a streamer's channel is flooded with abusive/hateful messages from bot accounts [59], or people use the raid mechanism to abuse a streamer [82]. Several streamers started a "#DayOffTwitch" campaign on Twitter to vent their dissatisfaction with the platform's slow and ineffective responses to these massive group attacks on marginalized streamers (e.g., black and LGBTQIA + streamers) [4]. Hundreds of bots joining the chatroom simultaneously disrupted the normal interaction among viewers with hateful message spam. Moreover, streamers and human moderators have to ban them with no end in sight.

While online harassment and content moderation are broadly investigated by HCI and CSCW scholars (e.g., [15, 17, 53, 64, 68, 72, 99, 105]), the new affordances of a platform provide new forms of violations that have never been caught before and disrupt the interaction experience of users. Attackers are creative and always abuse the features of a system to cause trouble to users on the platform, such as using the group voice chat on Discord to play porn to disrupt the voice discussion [48]. More broadly, as more marginalized and underrepresented groups also actively participate and diversify online communities, researchers have to be vigilant about exploring how the affordances of new technologies might be misused and abused [102].

In this study, we focus on coordinated group attacks in real time in live streaming communities, explore marginalized streamers' experience with *hate raids* on Twitch, and identify the challenges the communities face with potential design implications to cope with these attacks. We ask:

- RQ1: What are users' understandings and interpretations of hate raids?
- RQ2: What are the impacts of hate raids on live streaming communities?
- RQ3: What are the approaches and challenges to combat hate raids?

Through an analysis of scraped data on the Twitch subreddit (r/Twitch) about hate raids discussion, we contribute to understanding human-bot coordinated group attacks with real-time nature and towards marginalized users in live streaming communities. We clarify that the targets of hate raids are mainly marginalized streamers, but the hate raids discussion in this study is from all affected groups (e.g., marginalized streamers, general streamers, viewers, moderators, and streamers' friends, and some tool developers). We use affordances as a lens to explain how attackers leverage live streaming systems to conduct hate raids and the harm and trade-off framework to explain marginalized streamers' sufferings. We also propose the *moderation-by-design* concept (a concept suggesting that system design should always consider the potential abuse of such design and possible proactive and reactive responses to such abuse) to develop new features and moderation mechanisms and provide a list of recommendations and implications for stakeholders (platform, designers and developers, streamers, moderators, and viewers).

2 RELATED WORK

2.1 Harassment Towards Marginalized Groups and Content Moderation

Many scholars in HCI have explored online harassment in different contexts, such as thread comments [34], voice chat [48], and social VR [6]. In all these contexts, users experienced hate speech and harassment and suffered various intertwined harms: physical harm, such as self-injury and sexual abuse; emotional harm, such as depression and trauma; relational harm, such as damage to one's reputation and interpersonal relationships; and financial harm, such as loss of a digital asset or financial loss [88].

Women and LGBTQ communities are often targets of online harassment [20, 51] and experience more harmful behaviors than men, such as physical threats and sexual harassment [9], particularly

when online communities are dominated by men and performance is perceived as masculine, as in gaming communities [8]. Consequently, marginalized users are more likely to withdraw participation in online communities, or stay alone and anonymous with limited social signals to reveal their identities, such as using neutral avatars and avoiding using voice communication in gaming spaces [29, 55]. Marginalized users often lack of social support and experience emotional harms, such as anxiety and loneliness [73] and depression [65]. Their continued experience as "outsiders" urge cultural change in online environment [21].

Online harassment is either formed by an individual attacker or a group of attackers to initiate a hate campaign to attack a specific group, a marginalized group, or women in particular. Prior work does not clearly distinguish harassment by individual/random or group/organized. Organized harassment is less common than random attacks in many online communities and different from random attacks in several aspects. First, scalability makes general moderation strategies impossible; there is no effective strategy to stop the attack, such as educating and communicating with individual attackers in general online harassment [13]. Second, the intensity makes that no human moderator or algorithmic tool can effectively handle so much harmful content in a short time, and that the moderation action is less well planned and executed. For example, the notable *GamerGate* campaign on social media is a typical coordinated group attack on women in the video game industry in 2014 and 2015 [98]. Such large-scale online harassment is considered a semi-organized, pseudo-political movement on social networking sites [18]. The attackers are also less likely to be punished because it is challenging to detect their activities [19], curb the mix of human and semi-automated bots to spread manipulative content [107], and promptly remove high-volume postings in communities [18].

Content moderation refers to "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse" [36]. Many online communities more or less deploy content moderation that combines algorithmic tools and human moderators at the platform level; they also provide tools and features to end-users to customize the content (e.g., [46]). While algorithmic tools are used to detect and react to harmful content at scale, they also augment human moderators' capability of monitoring content and reviewing specific instances [31, 56]. Marginalized groups not only are easy to be targeted, but also experience disproportionate moderation [27, 33, 100]. For example, trans and black social media users often experience account and content removal regarding their marginalized identities and are limited to present in the public sphere [38].

A thread of research has explored the moderation mechanism from an individual entity perspective, both the user's and moderator's views. From the end-user's perspective, many scholars have tested and prototyped tools to mitigate harmful content to end-users and their communities, such as the creator-led comment-filter tool on YouTube [47], the language toxicity prediction and recommendation tool on Reddit [108], and evidence-capture tool on Facebook [99]. From the human moderator's perspective, a group of scholars has explored how moderators apply Twitch moderation tools to profile violators and manage viewers [11], how moderators configure and collaborate with Reddit Automod [45], and how interactive blurring tools mitigate moderators' exposure to harmful content [22].

Another thread has explored the early detection mechanism at scale (e.g., [78, 97]). For example, on Reddit, researchers have developed a predictive model with graphs, users, community, and text features, to create an early warning system for human moderators to prevent inter-community aggressive behaviors [54]; on YouTube, researchers have analyzed targeted videos' attributes to propose proactive moderation system to monitor hate attacks and mitigate their impact on content creators [69]. In this study, we aim to understand how live streaming users, particular marginalized

streamers, apply and evaluate the effectiveness of various moderation tools to deal with large-scale real-time coordinated group attacks.

2.2 Technological Affordances and Online Harassment

We use Norman's definition of affordances in HCI. Norman notes that affordances refer to the action possibility perceived by actors that an artifact offers for the action [79]. Technology can have distinct impacts on users based on their interpretation, intention, and knowledge [42]. In other words, users' online practices are not determined by technology, but by how they use it [79]. An affordance exists once users perceive a feature/function and the potential actions associated with it [74]. The same technology can be used differently among users [50, 80]. Thus, affordances can be shaped by both designers and users.

The Internet plays an essential role in shifting criminal opportunities from physical to virtual spaces [75]. Social networking sites can 'afford' the attack because they provide communication channels between victims and attackers and allow attackers to collect and disseminate information about victims [76]. Social media has several specific affordances regarding content distribution: persistence (easily recorded and archived), replicability (easily copied), scalability (easily shared), and searchability (easily accessed by others) [7].

Vitak et al. [102] summarize two types of affordances that social media platforms may afford online harassment. The increased content visibility makes harassment reach broad audiences and encourages potential harassers to join harassment activities. Anonymity and pseudonymity also encourage users to act more hostile since their real identities are hidden online, and they fear less loss of reputation [30, 93]. Therefore, under anonymity/pseudonymity, people would feel less responsible for their actions [95]. For example, research has shown that users who choose to be anonymous are more likely to post hateful comments; when an activity receives substantial hateful comments, it continues to receive such comments for a long time [110]. Marginalized users consider that anonymity and pseudonymity afford the safety for their community but also afford targeting and abusing from outside attackers [87].

Attackers use affordances of platforms to innovate ways to behave negatively [86]. For example, users often develop their understanding of the moderation system [44] and trick the algorithm with linguistic variations to avoid detection [15, 52]. Live streaming has some unique affordances, such as authenticity and synchronicity [13] and is initially designed to share and engage with demographically distant users to form communities. However, it is also used to facilitate sexual abuse in children, as attackers use it to broadcast sexual content and distribute it globally [41, 84]. Live chatroom for interaction also facilitates online harassment for the streamer as they stream with real-self and in real-time [101]. In this study, we focus on hate raids on Twitch, a form of synchronous communication in the chatroom with hateful messages flow. We use affordances as a lens to explain how hate raids abuse the affordances of live streaming systems to harm marginalized users.

2.3 Harassment and Content Moderation on Twitch

Live streaming as a novel media affords mass communication in the chatroom [39]. While streamers are broadcasting with various low- and high-fidelity equipment [24], viewers can simply register a pseudonymous account and send messages in the chat, and the streamer can read and respond orally to these messages. Oftentimes, the streamer would like to interact with the viewers to build communities [95] and promote prosocial behaviors in the chat [92]. Sometimes, attackers break the rules and start harassing the streamer/streaming content with toxic and hateful messages, even spamming these messages or emotes [85]. Twitch, a leading live streaming platform, is perceived as a masculine space dominated by white and male streamers [20]. Marginalized streamers (e.g.,

women and LGBTQ+) suffer various online harassment, such as sexually lewd comments and hate speech related to racism, sexism, homophobia, and transphobia; they have to manage their emotions and apply human moderators and tools to deal with harassment [101]. Hate raids as real-time coordinated group attacks exacerbate the aforementioned harassment with scalability and intensity by exposing the marginalized streamers in front of the camera. Limited resources and tools to handle such situations in time constraints can potentially intensify the harms experienced by marginalized streamers as they watch all these happening in real time.

Twitch applies a multi-level moderation system to combat harmful content on its platform. At the platform level, Twitch not only works hard with AI development to detect harmful content, but also hires employees to actively monitor all streaming activities. As Twitch's VP of trust and safety said, "We combine proactive detection, and a robust user reporting system with urgent escalation flows led by skilled human specialists to address incidents swiftly and accurately" [35]. At the community level, it allows each streamer to appoint human moderators and apply moderation tools, such as Twitch AutoMod and third-party tools [10], to manage the audience [106] and facilitate content moderation based on channel rules developed by the moderation team [12]. At the individual level, it provides settings for the viewer to filter the chat and block other viewers in the chat [2].

In this study, we focus on community-level moderation centered on streamers. Although a thread of research has explored moderation strategies to deal with harmful content [13], these strategies focus more on individual instances. Little is known about the generality of these strategies with regard to large-scale group attacks. We explore Twitch stakeholders' strategies and challenges to deal with hate raids.

3 METHODS

3.1 Data Collection

In this study, we collect comments on the Twitch subreddit (r/Twitch). Reddit is a large public online forum and divided into "subreddits", which are communities focused on specific topics and allowing users to join and post thread and leave comments. Posts and comments can be upvoted or downvoted, and can also receive awards. r/Twitch is the largest subreddit dedicated to Twitch, with approximately 1.2 million users at the time of data collection. This subreddit provides a neutral location for Twitch streamers, moderators, and viewers to share their experiences and seek advice on live streaming, such as how to set up live streaming equipment, what are the strategies to grow the viewership, how to deal with harassment and trolls in chatrooms, how to manage the viewership, and what are the updates about tools and policies from Twitch. We consider r/Twitch as the main source of data collection because (1) the first author has joined and followed the subreddit for more than two years; (2) hate raids mainly happened on Twitch, and r/Twitch is the initial and suitable place for Twitch users to discuss; (3) all timely discussion is archived with rich data types and samples, such as screenshot of hate raids on Twitch, resource to handle hate raids (e.g., external websites, shared files), streamers' complaint on other social media platforms.

We used the R 3.0.5 package "RedditExtractoR" to search for threads by keyword. We first use the terms "hate raids" and "follow bot" based on our observation of subreddit and news reports [59] and "hate attack" and "mob" in the literature review [69]. We read threads output based on each keyword. In this phase, we also tested and confirmed that the package could capture all variations with only a single term (e.g., hate raid, hate raided, hate raiding, and HATE RAIDS generated the same output). Next, we added new keywords based the output of first round reading, such as "massive bot" and "group attack." To iteratively read and compare outputs of these keywords, We also removed some keywords with their variations, such as "repeated message", "raid", and "mob", because these threads are about general spam and Twitch raid feature or irrelevant after reading

the threads, which do not capture the nature of hate raids. Hate raids as a new term unique to Twitch can cover most of the discussion if they are in the thread title and content. Finally, we used the keywords: "hate raid, follow bot, massive bot, hate attack, group attack". Then we merged the results, removed duplicates, and separated the results into two spreadsheets (Threads, Comments). The Threads contained the threads (the main posts), and the Comments contained the comments (replies to the main posts). The Threads spreadsheet includes the title (thread title) and text (detailed description of the topic). The Comments spreadsheet includes the text of the comment. We started data collection in January and collected 182 threads and 5392 comments in total till February 17th, 2022.

3.2 Data Analysis

We followed Fereday and Muir-Cochrane's six steps (codebook development, reliability test, initial theme identification, additional coding, theme identification, theme corroboration) and used a hybrid approach with inductive and deductive coding for the theme development [25]. First, three authors open-coded every thread on individual documents so that no author could influence another. If the threads (and later on the comments) contained a link to a different site, we followed the links and included their contents in the coding. Then, we shared the codes and discussed the similarities and differences between each rater's codes. By reconciling these codes, we generated a codebook with 20 meaningful codes plus two functional codes. Details are given in Appendix A.

Particularly, we coded the threads "not relevant (0)" if they are not clearly related to hate raids. We removed those irrelevant threads and the comments related to these threads consequently. We also added a code called "relevant but not in the list (22)" to apply to the Comments spreadsheet coding, just in case we missed something in the codebook development process. We finally kept 55 threads and 3,944 comments as final data for further analysis. Among the 55 threads, the earliest explicit description of hate raids was on 4/13/2021, but discussion exploded after a Public Service Announcement posted by r/Twitch moderators on 8/28/2021. Approximately 3% of the posts we analyzed were from before 4/13 (they were about related issues but not hate raids themselves), 3% were between 4/13 and 8/28/2021, and 94% were after 8/28. 58% of threads are created by streamers, 6% by moderators, 10% by viewers, and 26% by unknown. The most commented thread had 465 comments.

Second, two coders independently applied the codebook to a random sample of 100 comments in the Comments spreadsheet. We used this to calculate the inter-rater reliability with Cohen's Kappa. At first, the Kappa was .75, showing moderately significant agreement between the authors. However, the two authors met to discuss this inconsistency and realized that it was because some comments had multiple codes. After deciding to use only one code per comment, the section was recoded, and the reliability was recalculated. The Cohen's Kappa was .90, an almost perfect agreement. Third, we initially developed themes based on the codebook to primarily familiarize the topics. Fourth, two coders independently coded the rest of the Comments spreadsheet. Fifth, after coding, the three authors worked together to organize the codes into subcategories and high-level themes with the inclusion of code (22). Lastly, we iteratively grouped the categories into research questions with high quality examples for each category and adjust their fit.

4 RESULTS

Hate raids have been regarded as a prominent experience of marginalized (i.e., people of color, LGBTQ, female, and disabled) Twitch streamers. Many streamers use racial/identity tags to connect with their communities and promote themselves. However, the tags make the marginalized group prone to being attacked since the hate raiders *"target streams based on tags that were pro-LGBT or pro-equality"*. This finding supplements prior research about the Twitch tag that the tag system

increase the identity-based visibility but may introduce new ways for LGBTQIA+ streamers to be targeted [66]. In this section, if we explicitly know the comments from a affected group, such as streamer, viewer, and moderator, we mention it. If it is not clear, we use users in general. We use marginalized streamers and small streamers interchangeably to align with the literature and user discourse.

Hate raids take place not only in the live chatroom but also off the stream because viewers can participate in chat even if the stream is not live: *"These hate raids have been going to any offline channel do a bunch of stuff then report to Twitch that there's no modding chat."* Hate raids can also migrate to other platforms after the live streaming on Twitch. Discord is a social media platform with voice and video calls and text messaging. Users can form a community called "server" with a collection of categories and channels for users to join and interact. Since many streamers have group chats in Discord to have off-stream interaction with their followers, mass follow bots join the Discord servers of the steamers and post disturbing images (e.g., *"images of animal gore"*) and hateful words (e.g., *"pinging @everyone with a message containing targeted harassment & slurs"*).

4.1 RQ1: Twitch Users' Understandings and Interpretations of Hate Raids

4.1.1 Mass Bot Follows. Hate raids can be understood as mass bot follows. As a streamer experienced and summarized, *"If you're online, it clogs up your followers' alerts, which could last for minutes or hours until you pause the alerts or hide the source in your software. If done when you are offline, it basically means that any followers-only mode would be easily bypassed."*

Many other small streamers shared their experiences with the "hoss/host_XXXX" follow bots. These bots followed the streamers without posting or showing in the chat. As a small streamer said, *"None of those actually do anything to combat. They aren't in chat/chatting, just following and unfollowing",* as shown in Figure 1.

Some follow bots intentionally triggered the alert notification by "following and unfollowing", quoted from another streamer, *"They are constantly causing the following alert animation and sound to kick off, degrading the stream quality and making my viewers stop wanting to watch and stopping me wanting to stream."* By causing the notification sound, the "hoss" follow bots annoy the streamers and disrupt the streaming and viewing experience. Although there was a list of all the known over 1800 "hoss" follow bots shared among the streamers for their convenience to block the bots, some streamers argued that it was not helpful because *"the bots rename themselves often enough that by the time you've banned all known bots, there are new ones".*

4.1.2 Mass Hate Messages by Follow Bots or/and Human. Mass hate messages flooding the live chat in a short time can overwhelm the stream and ruin the community atmosphere. The messages can be sent by mass follow bots or humans. A user who believed that mass hate messages were caused by follow bots gave their definition of hate raids, *"A hate raid is when an account sets up a stream with several bots viewing and posting the same message over and over again raids another channel where the bots continue to post that same message over and over again. The message is usually something inflammatory and insulting."*

Since streamers have encountered different cases of hate raids, they had different understandings and definitions of hate raids. While some streamers received mass hate messages *"either shortly after (after a few seconds) or immediately after the bot has followed",* some streamers commented that not all follow bots would cause hate raids. Some users suggested that mass hate messages could be a mix of bots and human behaviors. A user defined hate raids as *"a mix between both raids and actual nasty people, and they target black people and people of the LBGTQIA+ community and basically spam a bunch of slurs and the n word".*

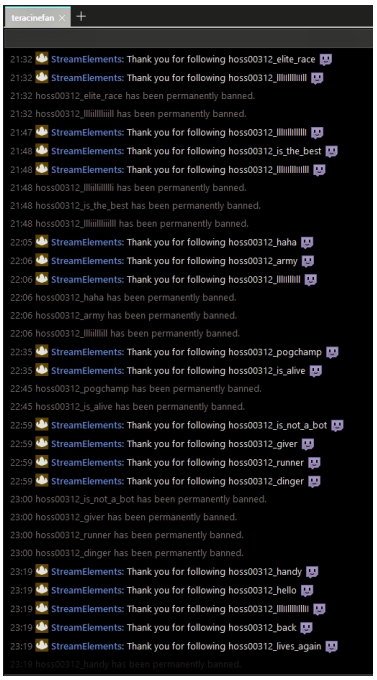

Fig. 1. A screenshot shared by a streamer at r/Twitch: the StreamElements (a moderation tool) notified new follows in the chatroom (in white color) and also banned hoss bots with notifications (in gray color). However, bots came back with new names, such as hoss00312_back and hoss00312_lives_again. Although there was no chat, the follow and ban notifications filled the chatroom.

There have been controversies and discussions on the "raids" features on Twitch. A user described hate raids as *"hundreds of bots suddenly following a streamer, and then posting heinous shit in chat. It doesn't have to be a Raid in Twitch language"*. There are some streamers with toxic community culture used the "raids" feature to pour their viewers into small streamers' channels and subvert their community sphere. Counterintuitively, while the literal explanation of hate raids is "a raid with hateful messages", hate raids do not have to involve the "raids" feature. According to a comment, *"Hate raids are raids usually containing bots that spam repeat messages in chat in large batches and groups. This does not use the actual raids feature to be performed."*

4.1.3 Human-engaged Streamer Entrapping. Some hate raiders attacked streamers as a group. They entrapped the streamers, induced them to violate Twitch's terms of services, and collected evidence to report the streamers to get them banned by Twitch. A viewer described a stream they watched that got a group attack:

> 1. Someone created a new account on her [the streamer's] Discord and posted disturbing/ graphic/ NSFW images. The images on the Discord screen showed up on the Twitch stream.
> 2. Someone on the Twitch stream said, 'Those images are against the Twitch terms of services. Since you streamed those images on Twitch, I have no choice but to report you.'
> 3. Someone created a new Twitch account meant to look like the streamer, but substituted a capital 'i' for a lowercase 'L', and then subscribed to her account, so it looked like

> 'AlissaSmith just subscribed to 'AlissaSmith' (example name, not the actual streamer). In the Twitch font, the capitol 'I' is almost indistinguishable from a lowercase 'L'.
> 4. Someone on another Twitch account said that the streamer was stupid for allowing the Discord images to show up, and called her a racial epithet (she is a person of color). While this was happening, the streamer got flustered and started crying, and one of the mods convinced her to end the stream and delete the VOD immediately. In retrospect, the attack was probably carried out by one person or a coordinated group of people. But at the time it seemed like just a bunch of weird random occurrences, and #2 and #3 didn't seem like obvious attacks.

Based on this viewer's description, the group attack was well organized and the role of every attacker was clear. They took advantage of Twitch's policies and its specific font to shut the stream, where the streamer had no chance to resist. Even after streamers carefully set all the moderation tools to proactively prevent similar incidents, many of them still fell into the trap designed by attackers and were banned by the platform. Even worse, the appealing process was also heavily delayed and alienated the streamers from their communities who could support them, as claimed by many victim streamers. A streamer shared their experience:

> I have all the nets in place. Follower-only chat, verified email, high level of moderation, removal of bots. I've been hit twice now. They know how to get around this stuff, if they want to take you out, they will. Each time I had a few guys come in and tell me they were about to get me ban. They spammed a link, I'm guessing to a porno site since that was what I was banned for. The link showed up as since I have links blocked. I deleted the messages with the blocked link but was still ban an hour later each time. I have sent twitch the screenshots of the chat, including DMs to my other social media accounts with 'that's what you get for banning me hahahaha' and 'Pssy ass btches like you deserve to die' from multiple random user accounts. Still waiting on my first appeal, let alone my second one.

According to this streamer, even after setting up all moderation tools with word filtering and deletion, when confronting a group of well-planned attackers, they were still so powerless and got banned. Furthermore, even when the streamer had tenable evidence handed to Twitch, the appealing process was still too slow to provide any practical assistance to the streamer.

4.1.4 Confusing Hate Raids For Human-engaged Trolls. There has been confusion between the term hate raids and group trolls. In this context, a troll is someone who is more mischievous than malicious. A streamer might think a hate raid is a group of trolls and not take it seriously, which could be harmful. A small streamer shared their experience of mistaking hate raids as trolls: *"In the beginning, I gave it time. I thought they may chill and it would be fun... but sadly, it started to be racist and annoying though I wasn't really actually annoyed just didn't feel like it was right to leave trolls enjoy freedom in my channel."* For new and small streamers like them, it was easy to mix the hate raids with group trolls and think that it is normal to have such hateful comments sent by the people whom they think are trolls, and they should learn to live with it to develop their channels and attract more viewers.

The confusion between trolls and hate raids can prevent streamers from protecting their stream and themselves timely. A mod shared the story of their streamer friend who was threatened with her life safety by hate raiders whom they mistakenly thought were trolls. Initially, they were told to ignore the mass of follow bots. However, after getting lots of hateful comments in the chat, they started to take the measures suggested by Twitch, such as adding bots, banning hateful words, and reporting the accounts. However, it did not seem helpful, and a few months later, the attackers not

only posted her personal information, her family's, and mods' through chat but also sent a package to her address, which meant they knew where she lived and messaged that they would rape her. *"It's extremely scary and things have been escalating for months, while we keep trying to ignore them and take the measures suggested by Twitch to stop them."* In this case, initially they thought that the attackers were just trolls, so they kept ignoring them, which escalated it to an unmanageable level: the personal information of her and the people around her was seriously compromised, and their safety received a huge threat.

4.2 RQ2: Impacts of Hate Raids on Live Streaming Communities

4.2.1 Streamers' Psychological Harm and Diminishing Community. Many streamers who had experienced hate raids reported that they were scared to stream or were not happy about it. Because they expected either follow bots or hate raids to happen while they were streaming, they were no longer enjoying streaming as before. As a streamer commented: *"Ever since these bots came in to full effect, every time I get a new follower (since I don't get many to begin with) I am IMMEDIATELY suspicious and start searching to see if they're a bot name. I hate doing it but damn it's just concerning now that any time I get a follow, I go into panic mode thinking I'm about to be ass blasted by hate spam."* Hate raids turned a new following from the most anticipated thing for a small streamer into something to be feared or even trauma. It also made streaming not about enjoying the games and interacting with viewers, but rather concerns about receiving attacks. As a result, streaming became emotionally burdensome and harmed mental health.

However, if attacked streamers stop streaming, their communities' engagement would decrease, harming their channels, which could severely hurt streamers who stream for a living or an essential source of happiness. A full-time streamer commented, *"My viewer count is dive bombing, since I am unable to make content in my primary game. Everything I have spent years building is crumbling, and there isn't a thing I can do to stop it, except feed these people, who have chosen to make it their full-time job to make my life a living hell."* Another user whose streamer friend took a few days off Twitch to avoid hate raids commented, *"The thing is streaming is her livelihood, she doesn't make much but it's enough to live off [of] currently. Not only that but it's the thing that makes her happy."* For transgender streamers, many mentioned they were not willing to turn their audio on during streaming because their voices can make them easier to get attacked, despite the fact that not having voice chat would negatively affect their viewership.

In addition to reducing or even withdrawing streaming, many streamers commented that they were concerned about their safety because some attackers had their personal information and could hurt them in real life. A streamer explained, *"He [the hate raider] said he was a killer and knew information he could only find from my social media account. Like can I go on my Twitch and see the IPs of people who looked on my account? I apologize for the hysteria but I am very much scared right now."* Such concern for personal safety could seriously influence streamers' daily lives. A user whose streamer friend had been attacked shared that their friend was not going out alone now and ordered food: *"Through a delivery gate so that no one can get to her in this manner,"* and was even considering moving: *"She lives alone, everyone is encouraging her to move, and she wants to. But apparently we can't break the rental contract over a stalker."*

Most viewers who witnessed hate raids would leave the stream because it affected viewing experiences. Additionally, some comments pointed out that viewers were concerned that they would be attacked if they stayed in the chatroom: *"Viewers leave their stream because they don't want to get doxed and don't want gross spammers sending them sexual whispers (some of the people getting aggressive rapey whispers are minors too which is even worse)."* Some viewers who shared the same identities with the streamer often left the room, fearing being attacked.

4.2.2 Users' Complaints and Sympathy Toward Twitch's Responses. Many complained that Twitch did too little to deal with hate raids. Some streamers commented that Twitch had few timely responses. A small streamer shared their experience of reporting attackers they encountered on Twitch: *"I reported it to Twitch with a list of the 100 or so names they used, but Twitch didn't do anything about it, they answered the ticket, but it was like a BS copy-paste kind of response (shrug). It's up to us to moderate ourselves really."*

In addition to the insufficient handling of hate raids reports, many users expressed their anger toward Twitch for its slowness in coming up with a systematic solution to the issue. A streamer commented, *"You gotta be kidding, people have been harassed for being part of a minority in twitch for a month, and twitch is only 'working on it', they should already have a solution, it's been a month already, people outside from twitch have come up with temporary solutions but no news from the company itself."* The streamer pointed out that while users promptly responded by developing temporary solutions to protect themselves and each other from hate raids, Twitch, which should have regulated attacks on its platform, was still so slow and could not provide any solutions or news.

Streamers who were unsatisfied with Twitch's response conducted the DayOffTwitch campaign and stopped streaming for one single day. While the campaign was designed to warn Twitch, it also influenced many Twitch streamers who were not originally attacked in the hate raids. Some small streamers took advantage of DayOffTwitch by streaming and got many more viewers than in the past because of less competition on that day. However, some other streamers lost followers or even were attacked for streaming. They complained about how the movement deviated from its original meaning of fighting against hate raids and became *"harassing people to protest harassment"*. Likewise, some streamers chose not to stream on the day just to avoid backlash.

However, while many users complained that Twitch was not effectively combating hate raids, some urged people not to expect too much from Twitch. On the one hand, there was a call for more patience with Twitch because it was difficult to devise a solution: *"They are working on it, this isn't as simple as 'incoming messages that look alike = block' They need to find a solution that doesn't block normal users, but to only block the bots, and that's not easy."* Also, since it takes time for Twitch to resolve issues, users should take some personal steps to protect themselves instead of solely relying on Twitch. A user wrote that, *"I'm not insisting Twitch can't do more, but I'm reluctant to place the blame solely on Twitch when we have personal moves we can make. It won't be perfect, but as I've said before, solutions are slow and hard to come by."* On the other hand, some users explained that it is hard for Twitch and every big game streaming platform to solve the hate raid issue, because *"every time you improve security, you get an influx of new 'talent' trying to break the system"*.

4.2.3 Weighing Other Platforms in Live Streaming Industry.

Users Intended to Leave Twitch but Facing Challenges. Some users suggested that leaving Twitch and joining other competitors could force Twitch to change: *"Change doesn't happen with overnight threats, it takes sustained throttling of their money avenues. If Twitch suffers, it suffers. Right now, they don't show for a minute that they actually care for their userbase. So if you want them to feel a hit, it needs to be a lasting impression."* Users believed that only when Twitch suffers long-term financial damage will it be forced to change the situation. However, switching to another platform was difficult for Twitch streamers, and the biggest issue was that their viewers might not be willing to move with them, so they could lose viewership. As a user wrote, *"95% of streamers aren't big enough and don't have the influence enough to bring their audience over to another platform. Even ninja lost almost half his viewers. Twitch knows it. We know it. That's why twitch is the way they are. They know most of us can't afford to start over."*

Users' Comparison Between Twitch and Its Competitors. In the discussion among users planning to leave Twitch and join its competitors, YouTube was frequently mentioned and compared to Twitch. Many users pointed out the advantages of YouTube, such as *"better copyright dispute system, larger platform and audience reach, better video and audio quality, unlimited VoD archival"* and *"better capacity to add the UI elements"* in terms of solving hate raids. Some users believed that *"YouTube can solve most of those problems easier than what twitch would have to do to get to bitrates and resolutions that compete with YouTube's"*. However, there were comments on why YouTube might not be a good choice for Twitch users. A user commented that streamers might experience similar harassment and attacks on other live streaming platforms as they have suffered on Twitch. YouTube is *"more strict in terms of monetization"* and has a *"more toxic and often a lot younger"* viewer community than Twitch. In addition to YouTube, some smaller platforms such as Trovo were also raised as options by some users: *"There is a competitor currently growing, it's called Trovo. It's pretty small but it has been growing at about the same pace Twitch did when it started."*

4.3 RQ3: Approaches and Challenges to Combat Hate Raids

Users have discussed social and technical approaches to combat hate raids, but social approaches are limited and are only discussed by a small group. Most discussions focused on moderation tools and whether they are effective. Not every suggested or existing tool can help solve these issues, but streamers were trying to compile a list of tools and settings to mitigate the impact of hate raids and follow bots. To this end, they were still looking for more effective moderation tools. The tools with descriptions are summarized in Appendix B.

4.3.1 Limited To No Discussion About Social Approaches. The social approach discussed mainly focused on how to support streamers with either "love raids" or moderation expertise. Some users suggested that streamers and viewers should rally to support those who were victims of hate raids. A user stated, *"I think it would be really cool if we could start a thread of streamers who could use a LOVE raid!"* Another user suggested that *"anyone who is struggling with trolls and haters in a dead chat, they should post in my discord to see who's awake and able to come help guide chat, deal with trolls (in a nice way), and restore your confidence"*. This would help streamers without established audiences since they are less likely to have fans or moderators to help them cope with an attack. This showcases the feeling of community that was widespread throughout these comments. While most people did not discuss this specifically and were not explicit about their support/care for the streamers affected, they showed how the users were concerned about hate raids and follow bots by giving suggestions.

4.3.2 Proactive Tools That Try to Prevent Attacks. Twitch chat settings, verification, IP bans, extension review, third-party tools, and improved control over raiding could be helpful before any attacks, by preventing malicious actors from accessing the chat. Verification, if enabled, would require all accounts to be verified before they were able to send messages. Ideally, this would stop attacks involving bots because each bot would have to be verified. A viewer pointed out that the main issue is the cost/benefit ratio for attackers:

> The question for an attacker is essentially: "is the time/effort for gathering accounts worth the attention I can receive from streamers?" This cycle currently fulfills itself because streamers at the moment cannot deal with the problem *until* it has happened, so even the act of starting a hate raid, even with the most effective of actions against such will significantly detriment a streamer and bring the attacker satisfaction.

Essentially, it is important to remember that attackers do this for their enjoyment. If Twitch makes it harder to create and verify bots, then the benefits to the attacker will be less than the cost

of the attack. This falls on Twitch to take care of, as streamers cannot prevent the creation of new bots.

The tools to which streamers have access also have problems. A streamer said, *"I've never seen this feature [requiring verification] work as intended because it's so easy to get around. In fact, you can use one single verified email to create thousands of bot accounts. How is that functional?"* More importantly, as a streamer said, *"Almost all streamers refuse to turn on"* the email verification option *"because they don't want to lose potential chatters."* Streamers did not want to have to choose between engagement and safety from attack.

Setting one's chat to only accept messages from those who meet certain standards could also significantly hinder attacks, especially with more stringent rules. Unfortunately, the same problem applies here as well. A viewer said, *"If I go to a channel, and it's followers only.... I leave."* For smaller streamers, who cannot afford to lose any potential viewers, this makes using strict chat settings untenable. If an attacker were IP banned, they would not be able to use alternative accounts to access Twitch, which would prevent what a streamer described as the game of *"Whack a HOSS [bot]."* This was something that many users wanted to see done more often, but there were two main problems. First, IPs are dynamic and can frequently change, so a completely random person could end up being banned while the attacker was still able to access the site. Additionally, a user pointed out that *"the bots are most certainly using VPNs and randomizing their IP addresses so blocking IPs would be entirely ineffective"*.

Reviewing extensions for security issues could help prevent people from falling victim to IP grabbers. Extensions add features like subtitles to streams but they often connect to external servers, which can pose a security risk. There was disagreement about whether or not this was a real problem worth spending time on. Those who believed it felt that it was another example of Twitch neglecting its responsibilities. One user said,*"Why can't Twitch verify extensions and only let creators use verified extensions? I mean they manually verify emotes right?"* However, there was much confusion about this point. For example, Twitch already does review extensions. One user pointed this out in return: *"And in fact that's what Twitch does. Alice&Slith had a really hard time getting their extension approved, which delayed their ARG for a few days/weeks."* This is also corroborated by Twitch's website [1], where the process for developing an extension clearly states that it must be reviewed before it is published.

Additionally, they pointed out that many worried about IP grabbers didn't really understand what an IP was for, and that there was not much cause for concern since every website someone interacts with can see their IP. Another user explained it by saying *"An IP address is sort of like a license plate number, its rather harmless information for people to have unless they intend to try to leverage everything they know about you against you."*

Of course, being able to reject individual raids would help streamers avoid that specific avenue of attack from any suspicious channels, even though most raids don't actually use the raids feature. Users felt that the options available, which were to either accept raids from everyone, only from friends, or turn off raids entirely, were too restrictive. A streamer said it was like *"a sledgehammer being used for a screwdrivers job"* as raiding was a fundamental way to network and expand one's audience. Like with verification and chat settings, streamers did not like sacrificing channel growth for safety. For marginalized streamers, their channels allow them to create a community where they can feel safe. Because of this, the prospect of losing engagement can be even worse.

4.3.3 Proactive Tools that Prevent the Audience From Being Exposed. Automod's word filter system is helpful if a channel has been attacked, as it prevents viewers from seeing messages that include specific words or phrases. Automod can be set to various levels of filtration, as well as banning

[1] https://dev.twitch.tv/docs/extensions/life-cycle/

or allowing specific words and phrases, so one streamer could ban all swearing while another could allow everything except for a word they did not like. Many users recommended it, but one major issue is that *"automod has a ton of issues with lgbtqia+ terms"*. Terms related to the LGBTQ community can be flagged as sexual content, which would mean that a streamer in the community could not take advantage of the filter's potential as it would ban discussion relevant to LGBTQ issues. In addition, attackers can easily bypass these filters. A black streamer commented that they *"used to get spammed with all kinds of symbols and hate"*, including *"getting called a 'rrigger'"*. As such, many users felt that Automod was an essential but inadequate tool to deal with these issues. Some third-party tools include moderation bots with lists of known hate raiders, which can be preemptively blocked. While there are some issues with this - chiefly, that this will always be out of date, as it takes time to add new accounts - this was something that many users appreciated, especially as these are *"simple to set up and just forget it"*.

4.3.4 Reactive Tools During an Attack. Chat settings and third-party tools can help mitigate the effects of an attack while it is occurring. If a hate raid happens, a streamer can either manually or with a third-party tool apply some settings that will stop the attackers from sending messages in the chat. Third-party moderation bots also allow streamers to set up a panic button, a command that executes multiple actions with one button press. A streamer stated, *"I have a Panic button (and an undo panic button) set up, so if for some reason I get some shit going down in chat, I press it, and everything is locked down... Mine will enable emote-only chat, sub-only chat, follow requirement of 5 or 10 min, clear chat, enable slow chat, disable alerts, etc., all with a single button press."* The benefit is clear, as it is a quick way to deal with an attack. As that streamer said, *"if you're prepared, you don't have to worry so much"*, which shows the peace of mind that this tool can provide.

An important difference between this and simply changing the chat settings from the beginning of the stream is that this does not harm engagement to the same degree. Rather than limiting participation from the beginning of the stream, this allows for a more flexible approach that only affects engagement during the attack. Afterward, the chat can be reverted to normal. This increased flexibility is something that users also wanted in regards to the raids feature itself, which shows that in several areas, Twitch hasn't given users as much control as they regard as necessary. A common theme was that while it was good that these third-party tools worked, it was a failure on Twitch's end that streamers had to rely on third parties. One user said, *"We should have the tools to protect ourselves and a panic button shouldn't be the only tool."* They were also frustrated that the panic button, which was universally considered an excellent tool, had been created by third parties when they believed this was something *"anyone could of done before"*.

4.3.5 Reactive Tools After an Attack. After an attack, bans would help prevent a recurrence from the same person. However, as explained earlier, regular bans and IP bans have drawbacks, so this will not solve the problem entirely. There are also third-party tools, such as serybot, that can identify and automatically ban known bot accounts. Additionally, some third-party tools give streamers the ability to review all followers within a specific time frame. If there is an attack, this allows them to ban all the bots at once.

On a more communal level, the victim of an attack could report the accounts involved to the developers of a third-party moderation tool, so that they could be added to the list of known attackers. Furthermore, many users wanted to implement multi-channel bans, which would allow communities of streamers to help each other. A person who is a moderator on multiple channels could ban an account, which would prevent that account from being able to access any of the channels that the person moderates. As a streamer described, *"This could be used to prevent users from harassing a group as a whole."* While this step does not help the individual streamer very much, since the attack is over by the time they can do this, it helps the streamer and the community overall

better prepared for the next attack. However, this is not very useful against attacks using bots, because it is easy for an attacker to get around bans, and they sometimes use names specifically taunting streamers who have banned them. As one user put it, *"You change your code, and attackers try different tactics. It's an arms race, really."*

5 DISCUSSION

In this study, we extend previous work on online harassment and content moderation and focus on the coordinated group attack in real time in live streaming communities. Hate raids as a human-bot coordinated group attack leverages the features of live streaming system to offend marginalize streamers with(out) violating rules. It initially targets marginalized streamers and can extend its targets to any streamers or user groups in live streaming communities, such as streamers who don't join the social media campaign and take advantage of the campaign. The attack pattern can also be generalized to any other platform, especially new platforms with many interactive elements but lack of moderation design. Marginalized streamers suffer from multiple harms, such as psychological harm, community loss, and safety threat. Marginalized streamers applies more technical than social approaches but can not sacrifice the engagement with high-level moderation like big streamers. The lack of effective tools and support to handle hate raids and the sufferings from it pile up marginalized streamers' complaint about and even contest against the platform. Such activities reflect the problems and challenges of the current moderation system design and urgently require new design approaches in the case of crisis management.

5.1 Affordances of Live Streaming Systems Facilitate Hate Raids

5.1.1 Features Abuse Without Violating Moderation Rules. Attackers utilize the Twitch features, initially designed to build a healthy live streaming environment and promote streamers, in negative ways to harm marginalized streamers. Streamers explicitly mention several features to exacerbate hate raids, making us reflect on the live streaming system design to mitigate feature abuse.

The identity tag mechanism increases the searchability of marginalized streamers, connects them to people who share the same identities, and promotes equality and their community [66]. However, it also provides potential attackers with opportunities to conduct hate raids and increases the scalability of attacks. The live streaming interface leads to asymmetric exposure between attackers and streamers [101, 111] and provides an environment for hate raids. On the one hand, the "live" affordances increase streamers' visibility to the public; on the other hand, they expose streamers' appearances and identities to potential attackers. Similarly, the text-based chatroom encourages viewers to engage while hiding attackers' identifiable information with pseudonymous usernames.

Prior work primarily emphasizes how attackers bypass the moderation system to keep breaking rules [15]. Similarly, the attackers keep generating new usernames to circumvent the block list developed by moderation teams and to send certain hateful words with variants in the chat. Differently, attackers leverage several features initially designed to facilitate live chat interaction. One prominent feature is the "follow" notification. Once a user clicks on it and starts following a streamer, the streamer will receive a notification from their side. Attackers trick the notification mechanism to keep "following/unfollowing" without breaking any rules to disrupt the interaction in the chat.

Attackers can also take advantage of the possible offline interaction on Twitch to circumvent active moderation and mode setting (e.g., follow-only mode). It is possible to access a streamer's chat even if the stream is not live, and devoted fans often maintain a small community there. Because offline chats are rarely active, most streamers do not moderate them. There were some anecdotes about attackers going into offline chats, sending hateful messages, and then reporting the streamers for having no moderation. Attackers even utilize Twitch's terms of service by entrapping streamers

to break rules and leverage the replicability of the Internet to secretly record live streaming screens, then report streamers to get them banned.

5.1.2 Algorithmic Confrontation Between Moderation Tools and Follow Bots. Moderation tools can actively ban bot accounts, but (1) the simple registration mechanism allows an email to generate multiple accounts, and even bots are allowed to register, and (2) each ban generates a notification in the chatroom. The ease of creating accounts provides fertile ground for attacks on marginalized groups [70]. The easy and quick generation of bots by attackers can use a simple algorithm to incessantly create new bots and send them into the chatroom. Though the tools can actively detect bots with similar usernames (e.g., hoss_xxx), the mass ban keeps generating notifications in the chat as bots keep joining in. Attackers leverage the simplistic account registration mechanism to confront the moderation algorithm. The synchronicity of the chatroom makes the algorithm confrontation feasible by generating flows of joining and banning notifications in the chatroom. Consequently, conversational messages can easily be lost in flooded notifications. Therefore, hate raids disrupt the conversation in the chatroom and cannot be stopped with available tools; there is no conversational resilience at all in this case [57].

5.1.3 Exploitation of the Platform Governance Structure. The imbalance between platform-driven and community-driven moderation can create space for potential hate groups to thrive [94]. Twitch's governance structure empowers its community moderation, nonetheless, making it harder to predict and detect hate raids. While proactive moderation tools are more at the platform level to detect video streaming violations, reactive moderation tools are more at the community level to combat chatroom violations. Community moderation like Twitch and Reddit empowers communities to develop their own rules and maintain their communities by themselves [91, 104]. While it gives users more power in moderation, various rules and norms challenge moderation at scale. No one algorithm/tool can handle multi-level platform governance. The streamer's chat could have enjoyable interaction at the beginning, which straightly passes the platform-level moderation. However, hate raids could suddenly happen in real-time with thousands of bots or with mixed humans and bots. Since hate raids circumvent proactive moderation tools, streamers must take immediate actions manually with their moderators. The large volume of bots and hateful messages overwhelms human labor. Additionally, bot-engaged hate raids might be a problem at the platform level, as they attacked a group of streamers at scale. However, human-engaged hate raids might only be a problem at the community level, depending on community rules and norms. Algorithmic design should consider the situated factors based on the governance structure, leaning a little toward community-level moderation.

5.2 Marginalized Streamers Endure Multi-level Harms and (Almost) Impossible Trade-offs Between Moderation and Participation

Prior work has explored marginalized streamers' emotional labor and management and different strategies to handle individual attacks with human moderators and tools [101]. This study extends this line of research by (1) highlighting other forms of harm caused by hate raids and how live streaming affordances and marginalization amplify these harms (2) and showing marginalized streamers' struggle to balance moderation and engagement, which is different from moderation strategies to handle individual attacks [13].

5.2.1 Multi-level Harms Caused by Hate Raids to Streamers. Streamers suffer multiple severe harms from hate raids. These harms are not just short-term, but long-lasting [103]. We align the harms with Scheuerman et al.'s harm framework [88] to explain different harms that marginalized streamers have suffered from hate raids.

Marginalized streamers experience mainly three harms: emotional harm, such as the panics of being hate raided and fear of restarting streaming; relational harm, such as streamer-viewer relationship disruption and concern about viewer engagement and community growth; and financial harm, such as viewership loss, community shrink, and subscription decrease. Although few streamers explicitly state physical harm, there is *potential* physical harm, such as safety threats with package delivery to their physical address. All these harms are intertwined. For example, safety threat increases psychological burdens with emotional labor and concerns; emotional harm, such as fear of streaming, and relation harm, such as viewership drop, can finally lead to a subscription decrease.

Several factors amplified the harms: (1) marginalized streamers are the targets, not the bystanders (perspective), (2) attackers highly intend to hurt them (intent), (3) marginalized streamers' harmful experience intensifies harm perception (experience), (4) hate raids in real-time are human-bot coordinated attacks (scale) and urgent to address with ineffective tools (urgency), (5) marginalized streamers are vulnerable (vulnerability), (6) hate raids can be textual with visual elements (e.g., emoji and memes) and with the video recording to entrap marginalized streamers (medium), and (7) hate raids happen in the public chatroom (sphere).

5.2.2 Trade-offs Between Participation and Moderation. Jiang et al. propose that content moderation is a series of trade-offs regarding moderation actions (e.g., excluding vs. organizing vs. norm-setting), styles (e.g., human vs. automated), philosophies (e.g., nurturing vs. punishing), and values (e.g., community identities) [49]. In this study, we align our findings with relevant trade-offs (actions, styles, and philosophies) to explain how hate raids are challenging to marginalized streamers' communities. Generally, the trade-offs are considered to deal with human-engaged hate raids. However, bot-engaged hate raids sometimes invalidate the trade-off framework and force marginalized streamers to accept the situation.

Regarding moderation actions, the synchronicity and bot-engaged hate raids basically make the trade-off of moderation actions invalid because (1) there is no way to exclude all bots, (2) there is no way to organize content as the instant notifications flowing in the chat, and (3) consequently, there is no way to have meaningful interaction in the chat to set a norm. Regarding moderation styles, the trade-off to deal with bot-engaged hate raids (e.g., mass follow bots with hate messages) is invalid because both humans and automation cannot deal with them effectively. Though automated moderation can constantly capture and ban bots, the algorithmic confrontation between tools and bots disables the interaction in the chat and also makes human labor powerless. The hybrid human and automated moderation might only deal with human-engaged hate raids to some extent. Regarding moderation philosophies, streamers are struggling to make a trade-off between punishing and nurturing. They have expressed concerns between the high-level moderation settings and viewer engagement. The settings (e.g., panic button) can at least mitigate part of hate raids. However, they usually work well with big streamers with a large user base without worrying about losing viewer engagement. Marginalized streamers, usually small streamers and the main target, might be forced to choose participation and community growth over moderation. Additionally, the bot-engaged hate raids might void the trade-off between level of activity and quality of contributions because the restricted moderation might not lower the bot activities at all and consequently increase viewers' contribution. Regarding moderation values, moderators and streamers do not have to make a trade-off because they usually share and maintain community identities.

5.3 Implications and Recommendations

Prior work shows that developers notice the flaws in system design to reactively recognize instances of harm to users, backtrack the causes, and fix the mistakes. Recently, Park et al. [81] developed a

prototype that can simulate different users and their interactions in either a positive or negative way, to some extent, can automatically identify harmful behaviors caused by the design so that developers can refine the design before deployment. Their research also sheds light on the moderation system design. While affordances indicate the perceived actions associated with the property of the features, the same feature can act in two opposite ways by regular users and potential attackers.

Combining our findings with prior work, we propose the *moderation-by-design* as a lens when designing new systems and moderation features. *Moderation-by-design* suggests that the mindset of system design should always consider the moderation elements, which is not only the design that facilitates cooperation, but also the mechanism that can potentially prevent abuse of such design. The mechanism should, from a socio-technical perspective, enable stakeholders to adapt and respond quickly through individual or collaborative actions, either proactively or reactively, and minimize the sacrifice of their existing experiences. Developers should consider the negative side when designing the moderation system and better understand the potential abuse of such system, though they have limited direct control over how their designs are enacted [50]. With this concept in mind, we provide the following recommendations. Before launching these features, developers should also simulate and test the potential abuse of these features. We propose the tool's design to focus on different stakholders' individual actions and collaboration to combat human-bot coordinated attacks. We clarify that these implications try to mitigate the impact of hate raids on participation; thus, some implications such as simply hiding "(un)following" notifications in the chatroom to avoid mass follow bots' impact are not listed because streamers lose the opportunities to interact with new viewers as well.

5.3.1 *Platform Governance With Communication Design.*

Better Channel to Engage with Marginalized Streamers. Prior work shows that protest users against platforms are more likely to be male and young users [62]; this study supplements prior work and shows that marginalized groups can also work as protest users. Twitch is the leading platform in live streaming industry with the possible performance of monopolistic practices [32], invisibly making streamers depend on it for daily needs. Their intention to leave Twitch and join competitors but unable to is in line with previous work that shows that the challenge is the concern of losing community connection [62]. This is a reminder to the platform to shift its focus on the big streamers who bring profits to the platform and to actively engage with marginalized streamers. For example, the platform can specifically add a channel to serve marginalized streamers and speed up the reporting and appeal processes.

Better Communication Between the Platform and Users. The mixed attitudes of the users towards Twitch indicate that Twitch needs better communication with its users to understand the problem and let its users know the necessary information. Although Twitch publishes the transparency report about moderation tools and settings [2], some settings are barely known from the user's perspective. For instance, Twitch developers have clearly stated in their UserVoice [2] communities that they have implemented IP Bans already, but the comments showed that this is not well-known by the public. The platform should consider better communicating its roles and actions to the public, at least to marginalized streamers, who always feel excluded and isolated.

5.3.2 *Implications for System Designer and Developers.*

Inclusive and Equitable Moderation Algorithm Design. Recent work also shows that different user groups consider toxicity differently; for instance, LGBTQ raters are more likely to annotate posts as toxic compared to random raters [33]. Similarly, mitigating online harassment needs to

[2] https://twitch.uservoice.com/

take marginalized users' needs into the platform and moderation system design [5, 90]. This is in line with Schoenebeck and Blackwell's notion about equality to equity for moderation system design [89]. This goal requires developers' input about design goals and rules and community members' values and needs, which require a strong developer and moderator/community member collaboration. However, prior research seemed to focus much on moderator-user interaction, moderator-bot interaction [13, 45] with little understanding of moderator-developer or community member-developer collaboration. For example, algorithm developers can work closely with LGBTQ+ streamers to update the terms in AutoMod to improve its efficiency and usability, in accordance with user-centered design methodologies throughout the design process [40, 50] and moderation system development [16]. A particular space to collect feedback from marginalized streamers could be considered.

Moderation System with Digital Forensics. Hate raid is not only simple online harassment, but also a kind of cybercrime, as Twitch sued attackers who conducted them [83]. There are many digital forensic tools to perform evidence analysis to identify potential crimes [43]. Existing moderation systems can also consider integrating some digital forensics technology to collect, preserve, extract, and report activities conducted by a user, for example, using digital forensic tools to investigate the streamer entrapping cases and tracing the whole process of attackers' behaviors instead of only relying on the image reported by attackers. This way can mitigate the reporting system's abuse and identify the attackers with a chain of evidence.

Engaging Third-party Developers as Ecosystem. Third-party developers are on the front line and usually encounter and react faster than the platform, as shown in this study. Third parties have already developed some tools to combat hate raids. We argue that third-party developers should be included in the moderation ecosystem [112] and that a mechanism should be implemented to facilitate professional and third-party developer collaboration, possibly by engaging third-party developers in the moderation algorithm design. Furthermore, certain parts of the algorithms should be efficiently utilized and modified by third-party developers.

5.3.3 Design to facilitate Streamer-Moderator Collaboration.

Tools to Increase the Visibility and Engagement of Moderators to Streamers. Streamers and moderators often work as a team to coordinate tasks and manage conflict [12, 14]. Sometimes, the streamer lacks active moderators in the chat to provide the necessary help [12], especially for new streamers [111]. We recommend tools to support how streamers can identify and need a resource from moderation expertise, for instance, a sidebar with a large available volunteer moderator list on Twitch homepage, showing volunteer moderators' preferences (channel, categories, streamer type, etc.). The list should be large enough to use massive human labor to temporarily join the chat to help the streamer. This perspective argues that streamers can use temporal massive human volunteer labor to combat massive human-bot coordinated attacks.

Tools to Support Moderation Team Posts in the Chatroom. Volunteer moderators create large commercial value for the platform, and the platform should show support for their voluntary work with more effective tools [61]. For example, TrollBuster, as a moderation tool to deal with real-time attacks on Twitter, allows a crisis response team to inundate the victim's Twitter feed with heart-warming and promising tweets to show emotional support to the victim [26]. Similarly, a tool should be designed to allow the moderation team to generate massive positive and encouraging messages in the chat when encountering human-engaged hate raids. However, bot-engaged hate raids are different scenarios that the moderation team cannot handle. In these scenarios, something like a conversational bot with positive message post settings should be considered.

5.3.4 *Design to Facilitate Streamer's Support Seeking and Viewers' Care Giving.*

Tools to Better Streamers' Support Seeking From the Same Identity Group. Prior work suggests that marginalized groups form their communities and safely disclose more about their experience and needs [37, 63]. Marginalized streamers might need more social support from their groups with the same identity on Twitch. Currently, they are sharing their experience granularly (e.g., on Reddit, Discord, Twitter, and other online communities). Twitch is considered a third space for streamer-viewer interaction. Possibly, it can also provide a space for steamers to network and seek different supports [101] from other streamers, such as instrumental, informational, and emotional support.

Tools to Stimulate Viewers' Positivity to Combat Human-engaged Hate Raids. Similar to the positivity generator idea by [3], some designs can be considered to promote counter-speech to combat human-engaged hate raids. Participating in massive live Twitch chat is less about self-expression and identification, but more about engaging in collective action consistently and continuously [28]. There are tools to promote counter speech from users when the streamer experiences hate speech [71] and to use CAPTCHA to verify human users and simultaneously stimulate positive emotions [92]. Tools to stimulate viewer engagement are helpful in dealing with human-engaged hate raids, usually just spamming text messages.

Tools to Crowdsource and Amplify Viewers' Positivity to Mitigate Bot-related Hate Raids. For bot-engaged or human-bot coordinated hate raids, we recommend a mechanism to facilitate and encourage passive users to use non-text-based communication [109] to impact the atmosphere in the chatroom. Therefore, a potential tool should be considered to support crowdsourcing practices of viewers in general, such as crowdsourcing moderation with up and down votes [58]. Similarly, designers can develop a feature to ensure encouraging messages on the top of the chatroom when the chatroom is full of bots with messages and notifications. This feature may require user-engaged communication tools to participate in content moderation when necessary. For example, a tool can add a stream overlay from the user's perspective to allow all viewers (with passive viewers) to vote [60] the love raids messages and stick them to the top of the chatroom to amplify the emotional intensity [67] so that the streamer can always see the positive and encouraging words on the top of the chatroom when there is no way to stop the hate raids with constantly flowing messages and notifications.

5.4 Limitation and Future Work

This study has several limitations. First, some quotes are from users in general with no clear roles. We are not sure whether they are streamers or viewers; thus, it might be hard to weigh their significance. Future work should collect data from different affected groups to enrich the depiction of hate raids. Second, the data collection is until February 17th, 2022. Since then, Twitch has been working on some solutions, such as giving steamers control of the "raids" feature [77]. There may be more discussion on effective solutions, though our findings suggest that hate raids can be totally irrelevant to the " raids" feature. Future research should try to explore the application of this feature and evaluate its effectiveness to supplement this study. Third, some themes can be explored further, such as how streamers and moderators collaborate to deal with hate raids. Lastly, we only focus on hate raids from the victim's view instead of the attacker's view. Though we menitoned attacks might migrate to other platforms, we know little about the attackers. Future research should explore how hate raiders form their communities and work with bots to attack other communities. This way, we can provide a holistic view about real-time group attacks.

6 CONCLUSION

In this study, we show hate raids as a new form of online harassment that targets marginalized streamers with both human- and bot-engaged attacks and leverages the affordances of live streaming systems to carry out these attacks. These attacks cause multiple severe harms to streamers and force streamers to accept situations with limited trade-offs. Marginalized streamers try more technical approaches rather than social ones, but lack effective tools. We propose moderation-by-design as a philosophy when designing future interactive systems to mitigate potential feature abuse and list suggestions and recommendations to users in live streaming communities.

ACKNOWLEDGMENTS

Thank Renkai Ma for the help with data collection. Thank Aashka Patel for the codebook development. This research was funded by by National Science Foundation (Award No. 1928627).

REFERENCES

[1] [n.d.]. Defining "Online Abuse": A Glossary of Terms. https://onlineharassmentfieldmanual.pen.org/defining-online-harassment-a-glossary-of-terms/

[2] 2021. H2 2021 Transparency Report. https://safety.twitch.tv/s/article/H2-2021-Transparency-Report?language=en_US

[3] Zahra Ashktorab and Jessica Vitak. 2016. Designing Cyberbullying Mitigation and Prevention Solutions through Participatory Design With Teenagers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 3895–3905. https://doi.org/10.1145/2858036.2858548

[4] Connor Bennett. 2021. What is 'A Day off Twitch'? Why streamers are striking to protest. https://www.dexerto.com/entertainment/what-is-a-day-off-twitch-why-streamers-are-striking-to-protest-1643209/

[5] Lindsay Blackwell, 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, CSCW (11 2017). https://doi.org/10.1145/3134659

[6] Lindsay Blackwell, Nicole Ellison, Natasha Elliott-Deflo, and Raz Schwartz. 2019. Harassment in Social Virtual Reality: Challenges for Platform Governance. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (11 2019), 1–25. https://doi.org/10.1145/3359202

[7] Danah Michele Boyd. 2008. *Taken out of context: American teen sociality in networked publics*. Ph.D. Dissertation. University of California, Berkeley. https://www.proquest.com/openview/9cc930ef134daf46c17434d2992e8251/1?pq-origsite=gscholar&cbl=18750

[8] Andrea Braithwaite. 2014. 'Seriously, get out': Feminists on the forums and the War(craft) on women. *New Media & Society* 16, 5 (8 2014), 703–718. https://doi.org/10.1177/1461444813489503

[9] Anna Brown. 2020. LGB online daters have positive experiences overall but face harassment. https://www.pewresearch.org/fact-tank/2020/04/09/lesbian-gay-and-bisexual-online-daters-report-positive-experiences-but-also-harassment/

[10] Jie Cai and Donghee Y. Wohn. 2019. Categorizing Live Streaming Moderation Tools: An Analysis of Twitch. *International Journal of Interactive Communication Systems and Technologies* 9, 2 (7 2019), 36–50. https://doi.org/10.4018/IJICST.2019070103

[11] Jie Cai and Donghee Yvette Wohn. 2021. After Violation But Before Sanction: Understanding Volunteer Moderators' Profiling Processes Toward Violators in Live Streaming Communities. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (10 2021), 1–25. https://doi.org/10.1145/3479554

[12] Jie Cai and Donghee Yvette Wohn. 2022. Coordination and Collaboration: How do Volunteer Moderators Work as a Team in Live Streaming Communities?. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14. https://doi.org/10.1145/3491102.3517628

[13] Jie Cai, Donghee Y. Wohn, and Mashael Almoqbel. 2021. Moderation Visibility: Mapping the Strategies of Volunteer Moderators in Live Streaming Micro Communities. In *Proceedings of ACM International Conference on Interactive Media Experiences*. 61–72. https://doi.org/10.1145/3452918.3458796

[14] Jie Cai and Donghee Yvette Wohn. 2023. Understanding Moderators' Conflict and Conflict Management Strategies with Streamers in Live Streaming Communities; Understanding Moderators' Conflict and Conflict Management Strategies with Streamers in Live Streaming Communities. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. ACM, Hamburg, Germany, 1–12. https://doi.org/10.1145/3544548.3580982

[15] Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016. #thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work and Social Computing*. 1201–1213. https://doi.org/10.1145/2818048.2819963

[16] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A Cross-Community Learning-based System to Assist Reddit Moderators. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30. https://doi.org/10.1145/3359276

[17] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The bag of communities: Identifying abusive behavior online with preexisting internet data. *Conference on Human Factors in Computing Systems - Proceedings* 2017-May (5 2017), 3175–3187. https://doi.org/10.1145/3025453.3026018

[18] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Hate is not Binary: Studying Abusive Behavior of \#GamerGate on Twitter. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. ACM, New York, NY, USA, 65–74. https://doi.org/10.1145/3078714.3078721

[19] Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Emiliano De Cristofaro, Gianluca Stringhini, and Athena Vakali. 2017. Measuring \#GamerGate: A Tale of Hate, Sexism, and Bullying. In *Proceedings of the 26th International Conference on World Wide Web Companion*. ACM Press, New York, New York, USA, 1285–1290. https://doi.org/10.1145/3041021.3053890

[20] Mia Consalvo. 2012. Confronting Toxic Gamer Culture: A Challenge for Feminist Game Studies Scholars. *Journal of Gender, New Media, and Technology* 1 (11 2012), 1–11. https://doi.org/10.7264/N33X84KH

[21] Amanda C.e Cote. 2017. "I Can Defend Myself": Women's Strategies for Coping with Harassment while Gaming Online. *Games and Culture* 12, 2 (2017), 136–155. https://doi.org/10.1177/1555412015587603

[22] Anubrata Das, Brandon Dang, and Matthew Lease. 2020. Fast, Accurate, and Healthier: Interactive Blurring Helps Moderators Reduce Exposure to Harmful Content. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8, 1 (2020), 33–42. https://ojs.aaai.org/index.php/HCOMP/article/view/7461

[23] Bryan Dosono and Bryan Semaan. 2019. Moderation Practices as Emotional Labor in Sustaining Online Communities. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3290605.3300372

[24] Ian Drosos and Philip J Guo. 2022. The Design Space of Livestreaming Equipment Setups: Tradeoffs, Challenges, and Opportunities. In *Designing Interactive Systems Conference*. ACM, New York, NY, USA, 835–848. https://doi.org/10.1145/3532106.3533489

[25] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods* 5, 1 (3 2006), 80–92. https://doi.org/10.1177/160940690600500107

[26] Michelle Ferrier and Nisha Garud-Patkar. 2018. TrollBusters: Fighting Online Harassment of Women Journalists. In *Mediating Misogyny*. Springer International Publishing, Cham, 311–332. https://doi.org/10.1007/978-3-319-72917-6{_}16

[27] Jessica L. Feuston, Alex S. Taylor, and Anne Marie Piper. 2020. Conformity of Eating Disorders through Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (5 2020), 1–28. https://doi.org/10.1145/3392845

[28] Colin Ford, Dan Gardner, Leah Elaine Horgan, Calvin Liu, a. m. Tsaasan, Bonnie Nardi, and Jordan Rickman. 2017. Chat Speed OP PogChamp: Practices of Coherence in Massive Twitch Chat. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, USA, 858–871. https://doi.org/10.1145/3027063.3052765

[29] Jesse Fox and Wai Yen Tang. 2017. Women's experiences with general and sexual harassment in online video games: Rumination, organizational responsiveness, withdrawal, and coping strategies. *New Media and Society* 19, 8 (3 2017), 1290–1307. https://doi.org/10.1177/1461444816635778

[30] Eric J. Friedman* and Paul Resnick. 2001. The Social Cost of Cheap Pseudonyms. *Journal of Economics <html_entglyph="@amp;" ascii="&"/> Management Strategy* 10, 2 (6 2001), 173–199. https://doi.org/10.1111/j.1430-9134.2001.00173.x

[31] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (7 2020), 1–5. https://doi.org/10.1177/2053951720943234

[32] Tarleton Gillespie, Patricia Aufderheide, Elinor Carmi, Ysabel Gerrard, Robert Gorwa, Ariadna Matamoros-Fernández, Sarah T. Roberts, Aram Sinnreich, and Sarah Myers West. 2020. Expanding the debate about content moderation: Scholarly research agendas for the coming policy debates. *Internet Policy Review* 9, 4 (2020), 1–29. https://doi.org/10.14763/2020.4.1512

[33] Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2

(11 2022), 1–28. https://doi.org/10.1145/3555088

[34] Nitesh Goyal, 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*. ACM, New York, NY, USA, 1–17. https://doi.org/10.1145/3491102.3517517

[35] Nathan Grayson. 2022. How Twitch took down the Buffalo shooter's stream faster than Facebook. https://www.washingtonpost.com/video-games/2022/05/20/twitch-buffalo-shooter-facebook-nypd-interview/

[36] James Grimmelmann. 2015. The Virtues of Moderation. *Yale Journal of Law and Technology* 17, 1 (2015), 68. https://digitalcommons.law.yale.edu/yjolt/vol17/iss1/2

[37] Oliver L. Haimson, Justin Buss, Zu Weinger, Denny L. Starks, Dykee Gorrell, and Briar Sweetbriar Baron. 2020. Trans Time: Safety, Privacy, and ContentWarnings on a Transgender-Specific Social Media Site. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (10 2020). https://doi.org/10.1145/3415195

[38] Oliver L. Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (10 2021), 1–35. https://doi.org/10.1145/3479610

[39] William A. Hamilton, Oliver Garretson, and Andruid Kerne. 2014. Streaming on twitch: Fostering participatory communities of play within live mixed media. In *Conference on Human Factors in Computing Systems - Proceedings*. ACM, New York, NY, USA, 1315–1324. https://doi.org/10.1145/2556288.2557048

[40] U. Hedestig and V. Kaptelinin. 2003. Facilitator's invisible expertise and supra-situational activities in a telelearning environment. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*. IEEE, 10. https://doi.org/10.1109/HICSS.2003.1173637

[41] Graeme Horsman. 2018. A forensic examination of the technical and legal challenges surrounding the investigation of child abuse on live streaming platforms: A case study on Periscope. *Journal of Information Security and Applications* 42 (10 2018), 107–117. https://doi.org/10.1016/j.jisa.2018.07.009

[42] Ian Hutchby. 2001. Technologies, texts and affordances. *Sociology* 35, 2 (5 2001), 441–456. https://doi.org/10.1017/S0038038501000219

[43] Abdul Rehman Javed, Waqas Ahmed, Mamoun Alazab, Zunera Jalil, Kashif Kifayat, and Thippa Reddy Gadekallu. 2022. A Comprehensive Survey on Computer Forensics: State-of-the-Art, Tools, Techniques, Challenges, and Future Directions. *IEEE Access* 10 (2022), 11065–11089. https://doi.org/10.1109/ACCESS.2022.3142508

[44] Shagun Jhaver, Darren Scott Appling, Eric Gilbert, and Amy Bruckman. 2019. "Did you suspect the post would be removed?": Understanding user reactions to content removals on reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (11 2019), 1–33. https://doi.org/10.1145/3359294

[45] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction* 26, 5 (2019), 35. https://doi.org/10.1145/3338243

[46] Shagun Jhaver, Christian Boylston, DIyi Yang, and Amy Bruckman. 2021. Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (10 2021), 1–30. https://doi.org/10.1145/3479525

[47] Shagun Jhaver, Quan Ze Chen, Detlef Knauss, and Amy X. Zhang. 2022. Designing Word Filter Tools for Creator-led Comment Moderation. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–21. https://doi.org/10.1145/3491102.3517505

[48] Jialun Jiang, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2019. Moderation challenges in voice-based online communities on discord. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (11 2019). https://doi.org/10.1145/3359157

[49] Jialun Aaron Jiang, Peipei Nie, Jed R Brubaker, and Casey Fiesler. 2022. A Trade-off-centered Framework of Content Moderation; A Trade-off-centered Framework of Content Moderation. *ACM Trans. Comput.-Hum. Interact.* (2022), 1–34. https://doi.org/10.1145/3534929

[50] Chris Jones, Lone Dirckinck-Holmfeld, and Berner Lindström. 2006. A relational, indirect, meso-level approach to CSCL design in the next decade. *International Journal of Computer-Supported Collaborative Learning* 1, 1 (3 2006), 35–56. https://doi.org/10.1007/s11412-006-6841-7

[51] Joann Keyton and Kathy Menzie. 2007. Sexually Harassing Messages: Decoding Workplace Conversation. *Communication Studies* 58, 1 (2 2007), 87–103. https://doi.org/10.1080/10510970601168756

[52] Soomin Kim, Changhoon Oh, Won Ik Cho, Donghoon Shin, Bongwon Suh, and Joonhwan Lee. 2021. Trkic G00gle: Why and How Users Game Translation Algorithms. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (10 2021), 1–24. https://doi.org/10.1145/3476085

[53] Yubo Kou. 2021. Punishment and Its Discontents: An Analysis of Permanent Ban in an Online Game Community. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (10 2021), 1–21. https://doi.org/10.1145/3476075

[54] Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community Interaction and Conflict on the Web. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. ACM Press, New York, New York, USA, 933–943. https://doi.org/10.1145/3178876.3186141

[55] Jeffrey H. Kuznekoff and Lindsey M. Rose. 2013. Communication in multiplayer gaming: Examining player responses to gender cues. *New Media and Society* 15, 4 (9 2013), 541–556. https://doi.org/10.1177/1461444812458271

[56] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q. Vera Liao, Yunfeng Zhang, and Chenhao Tan. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–18. https://doi.org/10.1145/3491102.3501999

[57] Charlotte Lambert, Ananya Rajagopal, and Eshwar Chandrasekharan. 2022. Conversational Resilience: Quantifying and Predicting Conversational Outcomes Following Adverse Events. *Proceedings of the International AAAI Conference on Web and Social Media* 16, 1 (2022), 548–559. https://ojs.aaai.org/index.php/ICWSM/article/view/19314

[58] Cliff Lampe and Paul Resnick. 2004. Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. 543–550. https://doi.org/10.1145/985692.985761

[59] Alexander Lee. 2022. 'Don't let it bother you, just continue streaming': Confessions of a Twitch streamer who received 'hate raids'. https://digiday.com/marketing/dont-let-it-bother-you-just-continue-streaming-confessions-of-a-twitch-streamer-and-victim-of-online-hate-raids/

[60] Pascal Lessel, Alexander Vielhauer, and Antonio Krüger. 2017. Expanding video game live-streams with enhanced communication channels: A case study. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, 1571–1576. https://doi.org/10.1145/3025453.3025708

[61] Hanlin Li, Brent Hecht, and Stevie Chancellor. 2022. Measuring the Monetary Value of Online Volunteer Work. In *Proceedings of the International AAAI Conference on Web and Social Media*. 596–606. https://ojs.aaai.org/index.php/ICWSM/article/view/19318

[62] Hanlin Li, Nicholas Vincent, Janice Tsai, Jofish Kaye, and Brent Hecht. 2019. How Do People Change Their Technology Use in Protest? *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (11 2019), 1–22. https://doi.org/10.1145/3359189

[63] Lingyuan Li, Kelsea Schulenberg, Guo Freeman, and Dane Acena. 2023. "We Cried on Each Other's Shoulders": How LGBTQ+ Individuals Experience Social Support in Social Virtual Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, Vol. 16. ACM, 1–16. https://doi.org/10.1145/3544548.3581530

[64] Na Li, Jie Cai, and Donghee Yvette Wohn. 2023. Ignoring As a Moderation Strategy for Volunteer Moderators on Twitch. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, Vol. 1. ACM, New York, NY, USA, 1–7. https://doi.org/10.1145/3544549.3585704

[65] Megan Lindsay, Jaime M. Booth, Jill T. Messing, and Jonel Thaller. 2016. Experiences of Online Harassment Among Emerging Adults. *Journal of Interpersonal Violence* 31, 19 (11 2016), 3174–3195. https://doi.org/10.1177/0886260515584344

[66] Jeremy Lopez and Guo Freeman. 2022. To Tag or Not To Tag: The Interplay of the Twitch Tag System and LGBTQIA+ Visibility in Live Streaming. *Proceedings of the 55th Hawaii International Conference on System Sciences* (1 2022). https://doi.org/10.24251/HICSS.2022.413

[67] Mufan Luo, Tiffany W. Hsu, Joon Sung Park, and Jeffrey T. Hancock. 2020. Emotional Amplification During Live-Streaming: Evidence from Comments During and After News Events. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (5 2020), 1–19. https://doi.org/10.1145/3392853

[68] Renkai Ma and Yubo Kou. 2021. "How advertiser-friendly is my video?": YouTuber's Socioeconomic Interactions with Algorithmic Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (10 2021), 1–25. https://doi.org/10.1145/3479573

[69] Enrico Mariconti, Guillermo Suarez-Tangil, Jeremy Blackburn, Emiliano De Cristofaro, Nicolas Kourtellis, Ilias Leontiadis, Jordi Luque Serrano, and Gianluca Stringhini. 2019. "You Know What to Do": Proactive Detection of YouTube Videos Targeted by Coordinated Hate Attacks. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (11 2019), 1–21. https://doi.org/10.1145/3359309

[70] Adrienne Massanari. 2017. #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media and Society* 19, 3 (2017), 329–346. https://doi.org/10.1177/1461444815608807

[71] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019*. 369–380. https://doi.org/10.13140/RG.2.2.31128.85765

[72] Brian McInnis, Leah Ajmani, Lu Sun, Yiwen Hou, Ziwen Zeng, and Steven P. Dow. 2021. Reporting the Community Beat: Practices for Moderating Online Discussion at a News Website. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (10 2021), 1–25. https://doi.org/10.1145/3476074

[73] Lavinia McLean and Mark D. Griffiths. 2019. Female Gamers' Experience of Online Harassment and Social Support in Online Gaming: A Qualitative Study. *International Journal of Mental Health and Addiction* 17, 4 (8 2019), 970–994. https://doi.org/10.1007/S11469-018-9962-0/FIGURES/1

[74] Joanne Meredith. 2017. Analysing technological affordances of online interactions using conversation analysis. *Journal of Pragmatics* 115 (7 2017), 42–55. https://doi.org/10.1016/j.pragma.2017.03.001

[75] Fernando Miró-Llinares and Asier Moneva. 2019. What about cyberspace (and cybercrime alongside it)? A reply to Farrell and Birks "Did cybercrime cause the crime drop?". *Crime Science* 8, 1 (12 2019), 12. https://doi.org/10.1186/s40163-019-0107-y

[76] Kimberly J Mitchell, David Finkelhor, Lisa M Jones, and Janis Wolak. 2010. Use of social networking sites in online sex crimes against minors: an examination of national incidence and means of utilization. *The Journal of adolescent health : official publication of the Society for Adolescent Medicine* 47, 2 (8 2010), 183–90. https://doi.org/10.1016/j.jadohealth.2010.01.007

[77] Ed Nightingale. 2022. Twitch streamers positive about new raid feature in response to harassment . https://www.eurogamer.net/twitch-streamers-positive-about-new-raid-feature-in-response-to-harassment

[78] Shirin Nilizadeh, François Labrèche, Alireza Sedighian, Ali Zand, José Fernandez, Christopher Kruegel, Gianluca Stringhini, and Giovanni Vigna. 2017. POISED: Spotting twitter spam off the beaten paths. In *Proceedings of the ACM Conference on Computer and Communications Security*. ACM, New York, NY, USA, 1159–1174. https://doi.org/10.1145/3133956.3134055

[79] Donald A. Norman. 1988. *The Psychology of Everyday Things*. Basic Books, New York. https://psycnet.apa.org/record/1988-97561-000

[80] Xinru Page, Andrew Capener, Spring Cullen, Tao Wang, Monica Garfield, and Pamela J. Wisniewski. 2022. Perceiving Affordances Differently: The Unintended Consequences When Young Autistic Adults Engage with Social Media. In *CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–21. https://doi.org/10.1145/3491102.3517596

[81] Joon Sung Park, Lindsay Popowski, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *The 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22)*. 1–18. https://doi.org/10.1145/3526113.3545616

[82] Ash Parrish. 2021. How to stop a hate raid. https://www.theverge.com/22633874/how-to-stop-a-hate-raid-twitch-safety-tools

[83] Ash Parrish. 2021. Twitch sues two alleged 'hate raiders'. https://www.theverge.com/2021/9/10/22666953/twitch-sues-alleged-hate-raiders-harassment-streamers

[84] Ethel Quayle. 2016. *Researching online child sexual exploitation and abuse: Are there links between online and offline vulnerabilities?* Technical Report. The University of Edinburgh, UK. 1–48 pages. www.globalkidsonline.net/sexual-

[85] Sarah Riddick and Rich Shivener. 2022. Affective Spamming on Twitch: Rhetorics of an Emote-Only Audience in a Presidential Inauguration Livestream. *Computers and Composition* 64 (6 2022), 102711. https://doi.org/10.1016/j.compcom.2022.102711

[86] D Robey, C Anderson, B Raymond Journal of the Association for, and undefined 2013. 2013. Information technology, materiality, and organizational change: A professional odyssey. *aisel.aisnet.org* 14, 7 (2013), 379–398. https://aisel.aisnet.org/jais/vol14/iss7/1/

[87] Morgan Klaus Scheuerman, Stacy M. Branham, and Foad Hamidi. 2018. Safe spaces and safe places: Unpacking technology-mediated experiences of safety and harm with transgender people. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (11 2018), 1–27. https://doi.org/10.1145/3274424

[88] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. 2021. A Framework of Severity for Harmful Content Online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (10 2021), 1–33. https://doi.org/10.1145/3479512

[89] Sarita Schoenebeck and Lindsay Blackwell. 2021. Reimagining Social Media Governance: Harm, Accountability, and Repair. *SSRN Electronic Journal* (7 2021). https://doi.org/10.2139/ssrn.3895779

[90] Kelsea Schulenberg, Lingyuan Li, Guo Freeman, Samaneh Zamanifard, and Nathan J Mcneese. 2023. Towards Leveraging AI-based Moderation to Address Emergent Harassment in Social Virtual Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, Vol. 1. ACM, 1–17. https://doi.org/10.1145/3544548.3581090

[91] Joseph Seering. 2020. Reconsidering Community Self-Moderation: The Role of Research in Supporting Community-Based Models for Online Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (10 2020). https://doi.org/10.1145/3415178

[92] Joseph Seering, Tianmi Fang, Luca Damasco, Mianhong Cherie Chen, Likang Sun, and Geoff Kaufman. 2019. Designing User Interface Elements to Improve the Quality and Civility of Discourse in Online Commenting Behaviors. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–14.

https://doi.org/10.1145/3290605.3300836

[93] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, New York, NY, USA, 111–125. https://doi.org/10.1145/2998181.2998277

[94] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21, 7 (7 2019), 1417–1443. https://doi.org/10.1177/1461444818821316

[95] Jeff T Sheng and Sanjay R Kairam. 2020. From Virtual Strangers to IRL Friends: Relationship Development in Livestreaming Communities on Twitch. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 34. https://doi.org/10.1145/3415165

[96] Miriah Steiger, Timir J. Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, 1–14. https://doi.org/10.1145/3411764.3445092

[97] Gianluca Stringhini, Pierre Mourlanne, Gregoire Jacob, Manuel Egele, Christopher Kruegel, and Giovanni Vigna. 2015. Evilcohort: Detecting communities of malicious accounts on online services. , 563–578 pages. https://www.usenix.org/conference/usenixsecurity15/technical-sessions/presentation/stringhini

[98] Keith Stuart. 2014. Zoe Quinn: 'All Gamergate has done is ruin people's lives'. https://www.theguardian.com/technology/2014/dec/03/zoe-quinn-gamergate-interview

[99] Sharifa Sultana, Mitrasree Deb, Ananya Bhattacharjee, Shaid Hasan, S.M.Raihanul Alam, Trishna Chakraborty, Prianka Roy, Samira Fairuz Ahmed, Aparna Moitra, M. Ashraful Amin, A.K.M. Najmul Islam, and Syed Ishtiaque Ahmed. 2021. 'Unmochon': A Tool to Combat Online Sexual Harassment over Facebook Messenger. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–18. https://doi.org/10.1145/3411764.3445154

[100] Hibby Thach, Samuel Mayworm, Daniel Delmonaco, and Oliver Haimson. 2022. (In)visible moderation: A digital ethnography of marginalized users and content moderation on Twitch and Reddit. *New Media & Society* (7 2022), 1–22. https://doi.org/10.1177/14614448221109804

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

[102] Jessica Vitak, Kalyani Chadha, Linda Steiner, and Zahra Ashktorab. 2017. Identifying women's experiences with and strategies for mitigating negative effects of online harassment. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*. 1231–1245. https://doi.org/10.1145/2998181.2998337

[103] Ashley Marie Walker and Michael A. 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*. ACM, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376497

[104] Donghee Yvette Wohn. 2019. Volunteer Moderators in Twitch Micro Communities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300390

[105] Donghee Yvette Wohn, Casey Fiesler, Libby Hemphill, Munmun De Choudhury, and J. Nathan Matias. 2017. How to handle online risks? Discussing content curation and moderation in social media. *Conference on Human Factors in Computing Systems - Proceedings* Part F127655 (5 2017), 1271–1276. https://doi.org/10.1145/3027063.3051141

[106] Donghee Yvette Wohn and Guo Freeman. 2020. Audience Management Practices of Live Streamers on Twitch. *IMX 2020 - Proceedings of the 2020 ACM International Conference on Interactive Media Experiences* (6 2020), 106–116. https://doi.org/10.1145/3391614.3393653

[107] Samuel C Woolley. 2022. Digital Propaganda: The Power of Influencers. *Journal of Democracy* 33, 3 (2022), 115–129. https://muse.jhu.edu/article/860232

[108] Austin P. Wright, Omar Shaikh, Haekyu Park, Will Epperson, Muhammed Ahmed, Stephane Pinel, Duen Horng (polo) Chau, and Diyi Yang. 2021. Recast: Enabling User Recourse and Interpretability of Toxicity Detection Models with Interactive Visualization. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (4 2021). https://doi.org/10.1145/3449280

[109] Mu Xia, Yun Huang, Wenjing Duan, and Andrew B. Whinston. 2009. Ballot box communication in online communities. *Commun. ACM* 52, 9 (9 2009), 138–142. https://doi.org/10.1145/1562164.1562199

[110] Savvas Zannettou, Mai Elsherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. 2020. Measuring and Characterizing Hate Speech on News Websites. In *12th ACM Conference on Web Science*. ACM, New York, NY, USA, 125–134. https://doi.org/10.1145/3394231.3397902

[111] Yingfan Zhou and Rosta Farzan. 2021. Designing to Stop Live Streaming Cyberbullying. In *Proceedings of the 10th International Conference on Communities & Technologies - Wicked Problems in the Age of Tech*. ACM, New York, NY,

USA, 138–150. https://doi.org/10.1145/3461564.3461574
[112] Ethan Zuckerman. 2021. Why study media ecosystems? *Information Communication and Society* 24, 10 (2021), 1495–1513. https://doi.org/10.1080/1369118X.2021.1942513

A CODEBOOK

- (0) Not relevant
- (1) Twitch sues hate raiders
- (2) Follow bots
- (3) Solutions to hate raid
- (4) Use HR movement to self promote
- (5) DayOffTwitch useless
- (6) Hate raider community outside Twitch
- (7) Recommended tools to combat hate raid
- (8) Suggestions to Twitch to combat hate raid
- (9) Ineffective tools
- (10) Official Twitch tools to combat hate raid
- (11) Follow bots grab IPs
- (12) Definition of hate raid
- (13) Small streamers targeted
- (14) Call for cultural change among Twitch users
- (15) Streamers share stories
- (16) Speculation Twitch users leave to join competitors
- (17) Ghost viewers
- (18) People expecting too much from Twitch
- (19) Minority streamers being targeted
- (20) Psychological impact of hate raids
- (21) Engagement impact of hate raids
- (22) Relevant but not in list

B TOOLS DESCRIPTIONS WITH USERS' OPINIONS

Table 1. Tools with Description and User's Attitudes

Tool/Feature	Description	What Supporters Believe	What Opponents Believe
Automod	Automod is provided by Twitch to prevent harmful chat messages from being seen by other viewers and includes levels of moderation and a customizable word block list.	• Built in tool - easy to set up • Automatically catch harmful messages • Moderator has to manually ban/mute, so mistakes are easily resolved	• Can't automatically ban/mute, so streamer has to be actively involved • Word filters can be easily circumvented
Twitch Chat Settings	Twitch chat settings modify who is allowed to send messages in chat and what those messages are like - for example, a streamer or moderator can set their chat to only accept messages from those who are followers.	• Built in features - easy to set up • Can prevent attackers from sending harmful messages	• Has to be manually modified for different situations • Safer (more restrictive) chat settings can decrease engagement
Verification	Twitch allows streamers to allow only those who have verified their accounts (by email, phone, or both) to send messages.	• Attackers have to spend more time verifying bot accounts • If one account linked to an email is banned from a channel, all accounts linked to that email are as well	• It is easy to create many emails to verify bots • Possible (though more difficult) to do same with phone numbers • Most importantly, these decrease engagement
Reviewing Extensions	Twitch allows channels to use extensions to enhance their stream (such as captions). Some extensions contact external servers , which could allow an attacker to log a viewer's IP address. Because of this, some users wanted Twitch to manually review every extension for security issues.	• Reviewing these extensions would reduce the chance of personal information being leaked	• Extensions are already reviewed before they are released • Everything you connect to can see your IP, and it isn't very dangerous, so this is a waste of resources
IP bans	While banning individual bot accounts is pointless, banning IP addresses associated with a set of bot accounts would take them all offline at once.	• More efficient and less frustrating than banning individual bots • Twitch already sometimes IP bans reported accounts	• IPs are dynamic and change, so the bot could one day receive a new, unbanned IP • If these accounts use VPNs, they can easily change their IP address
Increasing Streamer Control Over Raids	Raiding is an important feature on Twitch, even if it is sometimes misused for hate raids (though most hate raids do not actually utilize the raid feature). Streamers can either reject all raids, allow only from friends, or allow from everyone.	• Restricting or disabling raids hinders channel growth • Accepting all raids leaves streamers vulnerable • Allowing streamers to accept or reject individual raids would avoid both of these issues	• This will have a small impact on hate raids
Third Party Tools	Third party tools are bots created to help streamers and moderators moderate their streams and provide other services unrelated to moderation. These are not made by Twitch, but the most popular ones have hundreds of thousands of users.	• Gives streamers more flexibility and control • Gives streamers peace of mind	• Many commenters were frustrated that third parties were working on problems that they believed Twitch was ignoring

Received January 2023; revised April 2023; accepted May 2023