Moderation Visibility: Mapping the Strategies of Volunteer Moderators in Live Streaming Micro Communities

Jie Cai
New Jersey Institute of Technology
USA
jie.cai@njit.edu

Donghee Yvette Wohn
New Jersey Institute of Technology
USA
yvettewohn@gmail.com

Mashael Almoqbel
New Jersey Institute of Technology
USA
ma735@njit.edu

ABSTRACT

Volunteer moderators actively engage in online content management, such as removing toxic content and sanctioning anti-normative behaviors in user-governed communities. The synchronicity and ephemerality of live-streaming communities pose unique moderation challenges. Based on interviews with 21 volunteer moderators on Twitch, we mapped out 13 moderation strategies and presented them in relation to the bad act, enabling us to categorize from proactive and reactive perspectives and identify communicative and technical interventions. We found that the act of moderation involves highly visible and performative activities in the chat and invisible activities involving coordination and sanction. The juxtaposition of real-time individual decision-making with collaborative discussions and the dual nature of visible and invisible activities of moderators provide a unique lens into a role that relies heavily on both the social and technical. We also discuss how the affordances of live-streaming contribute to these unique activities.

CCS CONCEPTS

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

KEYWORDS

live streaming; content moderation; workflow; volunteer moderators; moderation strategies

ACM Reference Format:
Jie Cai, Donghee Yvette Wohn, and Mashael Almoqbel. 2021. Moderation Visibility: Mapping the Strategies of Volunteer Moderators in Live Streaming Micro Communities. In *ACM International Conference on Interactive Media Experiences (IMX '21), June 21–23, 2021, Virtual Event, NY, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3452918.3458796

1 INTRODUCTION

Online communities provide the opportunity for millions of users to express themselves and exchange information. Freedom of speech leads to complicated challenges for these online spaces, such as hate speech and harassment. Prior literature has discussed the management of negative content from various perspectives, such as moderation techniques [57], algorithms [1], level of discourse [20], commercial labor [46], users [31, 43], policy [21], law [35], and so forth, but it is still challenging to effectively moderate these contents as the communities evolve. Live streaming, as a unique social medium with high-fidelity computer graphics and video and low fidelity text-based communication [25], is a rapidly growing industry, estimated to reach 70.5 billion USD by the year 2021 [40], and also suffers from the toxic textual content. In this study, we extend previous research by focusing on the volunteer moderators' moderation practices in live-streaming communities.

Recent work of volunteer moderators and moderation mainly focuses on user-governed platforms such as Wikipedia [11, 33], and Reddit [9, 16, 29]. Twitch, as a user-moderated, live-streaming community, is similar in some governance aspects to other online communities such as Reddit, which is a self-reliant community [29], and Facebook Group, which provides multiparty interactions [51]. However, the interactivity of live streaming makes it different from other social platforms in mainly three aspects: 1) the large volume of messages generated and posted in a short time, 2) the flow speed of these messages in the chat, and 3) the limited time for the moderator to remedy harmful situations. These unique affordances may exacerbate moderation challenges.

Prior research about live streaming mainly focuses on the entertainment elements and explores the streamer or the viewer motives (e.g., [3, 6, 17, 25, 48]), the streamer-viewer interaction (e.g., [12, 36, 60]), and streamers' regulation of the broadcast [56, 59]. Only limited research has examined the negative aspects related to human moderators and content moderation. Some research explores moderators' emotional tolls [58], and the moderator selection process [52, 58], but little work explores the strategies that moderators utilize and their mental models of decision-making. This study contributes a moderator-centered perspective to the growing body of literature on volunteer moderation, considering how moderators develop and apply these strategies in live streaming communities, where broadcasters showing their face have heightened vulnerability and as real-time interaction between broadcasters and viewers make harassment difficult to avoid and handle.

Through 21 semi-structured interviews with Twitch moderators, this work has mainly twofold contribution: 1) We highlight the visible activities that volunteer moderators perform during the moderation process, which has been previously described as activities that usually happen behind the scene and lack transparency; 2) We develop a diagram to show the workflow of moderation with an emphasis on the communicative components in 'live' moderation systems. We discuss how the interactivity of live stream facilitates the moderation visibility and how the synchronicity enhances the graduated moderation and amplifies the violator's voice

IMX '21, June 21–23, 2021, Virtual Event, NY, USA

ACM ISBN 978-1-4503-8389-9/21/06…$15.00
https://doi.org/10.1145/3452918.3458796

in the workflow. Given the growing interest in using algorithmic methods to detect negativity [37], automate moderation [8], and build moderation tools [4, 50], these results provide further insight into the work of volunteer human moderators, offering potential directions into future research on the socio-technical interaction that takes place in live streaming communities as well as the design of these spaces.

2 RELATED WORK

2.1 Community Moderation, Human Moderator, Moderation System

Content moderation refers to *"the organized practice of screening user-generated content posted to internet sites, social media, and other online outlets, in order to determine the appropriateness of the content for a given site, locality, or jurisdiction"* [47], happening from online forums to social media platforms (e.g., [13, 43, 61]). Social media platforms often apply algorithms to detect misbehavior at scale [22]. They also employ a large group of commercial content moderators or freelancers who work on contract with them [46] to supplement these algorithms. In addition, platforms rely on users' reports who flag the potentially offensive content and then ask the moderator to review and remove the content manually [14, 19]. Users can also apply tools such as 'Blocklist' on Twitter to block harassers [31].

Different from most social media platforms that handle moderation within, user-governed communities such as Wikipedia and Reddit rely on volunteer moderators who are given limited administrative power to remove unacceptable content and ban violators [41]. These moderators are either selected from among the users who are most actively involved in the community and who are invested in its success [52, 58], or self-appointed, depending on the platform. Those who become moderators due to their high level of activity usually have a better understanding of the values and expectations of the communities [58].

The nature of the role of volunteer moderators can be social and communicative in user-governed communities [38]. Current work has explored the relationship between moderators' actions and end-users' responses [53], and has discussed how human moderators apply moderation tools to curate content [29], collaborate with other moderators [42], and identify norms violations [7, 9]. Moreover, Fiesler et al. found that moderators, though are willing to model good behaviors, prohibit bad behaviors more [16], indicating that moderators are likely to communicate but not very frequently.

Many existing moderation systems are relying on either algorithms or human moderators that lack transparency. The algorithmic content moderation at scale suffers from opacity without explanation after content removal [22, 23]. Current work considers commercial content moderators as the *"hidden labor"* behind the scene [46], and their work is hard to be seen by the end-users [43]. Although the combination of algorithms and commercial moderators can curtail harmful content, the current moderation system on social media platforms can cause some frustration due to its black-boxed nature; for example, content removal without any explanation, appeals processes that seem to go nowhere, and minimal opportunities for users to interact directly with the administrators [43]. The challenges of the current moderation systems of social media provide an opportunity for a new moderation system that can educate and engage users at the same time [30, 43].

2.2 Moderation in Live Streaming and on Twitch

While much previous work has focused on moderation in asynchronous online communities and social media platforms, very little is known about human moderation in synchronous online communities with live interaction among users in a timely manner. Recent research about moderation in live streaming focused on the motivation of being a moderator [58], how moderators engage with their communities [51], and categorizing moderation tools [4]. However, there has been relatively less discussion about the on-the-ground moderation practices of volunteer moderators– namely what kinds of strategies they use, and how these strategies work together during the moderation process, a gap which the present work aims to fill. Thus, we ask:

- **RQ:** *What is the workflow of volunteer moderators in live streaming communities? Specifically, what are the strategies and how are they connected?*

Twitch has become one of the global leading live-streaming platforms [18] and is interesting from a moderation perspective because of its platform design and affordances. Twitch users form micro-communities (channels) around the streamer [58] that each operates under different rules, has different audiences, and is responsible for the moderation of its chatroom. There are over a million streamers on Twitch, yet none of the channels are exactly the same; what may be acceptable in one channel may not apply to the other. It has a different conversational structure (messages appear chronologically) compared with Twitter or Reddit, which applies threaded conversations [49]. Additionally, the nearly-synchronous conversation in the chatroom requires more immediate attention from moderators [52].

To handle these chat messages, Twitch also employs both algorithms and human moderators, although it continues to change its structure. Until 2019, the company employed commercial moderators who mostly handle inappropriate broadcasting content that has been reported by users [45], but it primarily relies on volunteers to manage the chatrooms. It also has a moderation tool called AutoMod that relies on algorithms to help streamers moderate their chatrooms. In addition, it has an open access Application Programming Interface (API) to allow third-party tools. The moderation tool on Twitch could effectively discourage spam, and specific types of negative behaviors [50]. However, the quality and functionality of bots still pose some social and practical challenges [13, 39]. Streamers also employ the help of volunteer moderators. The volunteer moderators on Twitch are appointed by the streamer and help the streamer manage the chat content. When a viewer comes to the chat and starts typing, the chat rules will pop out. The viewer has to click "OK" to acknowledge the rules and to start chatting. The viewer can mention anyone in the public chat using "@" or can start a private conversation via the Whisper function under the user's profile. The moderators have to track a high volume of fast-moving messages, identify the negative comments, and take actions within a limited time. It is still not immediately clear how they moderate in live-streaming communities. Twitch offers various badges to

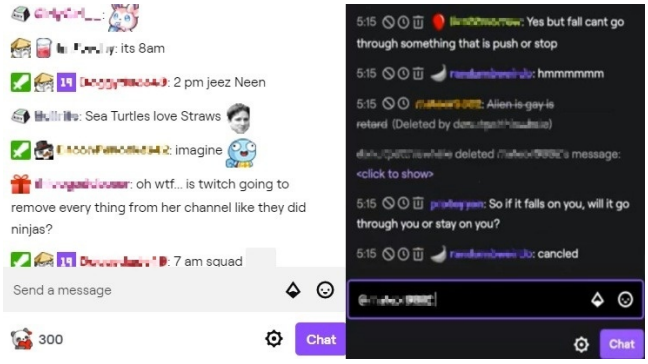

Figure 1: Twitch Chatroom Interface from Viewers' View (Left, colorful usernames with different badges to indicate status) and Moderators' View (Right, Shortcuts of "Ban""Timeout""Delete" are visible next to the usernames)

viewers to represent their status and indicate their contribution to the communities and micro-communities. Volunteer moderators have a special badge, a small icon containing a white sword with a green background. Figure 1 shows the interface of Twitch chatroom from both the moderator's and viewer's perspective.

3 METHODS

3.1 Participant Recruitment

The project and interview protocol were reviewed and approved by the Institutional Review Board (IRB). We recruited volunteer moderators from Twitch in four ways. First, we used the official Twitter account of our lab to post a recruitment message, and at the same time, searched for moderators with search terms such as "Twitch mod" and "moderator on Twitch". If someone was interested, they could send us messages through Direct Message (a message feature of Twitter), or if we found someone, a recruitment message would be sent through Direct Message. We obtained 10 moderators through Twitter. Second, private Twitch accounts were used to reach out to four moderators by directly messaging active moderators in random channels through Whisper (a message feature of Twitch). Third, two moderators who were acquaintances or friends of acquaintances of the researchers were recruited. Last, we reached out to five moderators through the recommendation of streamers that were interviewed for another project. Each of the 21 moderators received a $20 Amazon gift card for their voluntary participation.

3.2 Interview

Most interviews were conducted through Discord (a VoIP communication application) with a length between 40 and 60 minutes. During the interview, we first asked general questions about moderation experience such as *"Who are you a mod for?"* and *"How long have you been a mod?"*. Then we asked main questions related to our research questions such as *"How do you know how to mod?"" Do you have any prior experience?" "How do you decide what is appropriate or not?"* and *"How do you deal with toxicity and harassment?"* with the following questions like *"How do you decide when to ban, versus time out or ignore?"*. In the end, we asked them about anything that we did not mention, and they would like to share. We then closed the interview with a brief demographic indication (age, race, and gender). The beginning and end parts of the interview protocol are partially summarized in Table 1.

In order to have a big picture of moderation strategies and their relationship, we used thematic analysis [2] to code answers into concepts and group the relevant concepts into themes. After completing the semi-structured interviews and transcriptions, we first pasted all interview questions and corresponding answers into a spreadsheet, where all researchers went through the content of each transcript and became familiar with their content. To obtain a clear picture of themes, we grouped all the interview questions and related answers and perceptively put them under the two research questions. Second, an open coding approach was used iteratively; each researcher coded a group of interview questions and presented codes to each other in regular face-to-face calibration meetings, followed by a group discussion to clarify the consistency and accuracy. For example, the high-level category *"live explanation"* contained subcategories such as *"offering help and providing suggestions", "asking the viewer to leave",* and *"warning with prohibition"* with more detailed codes such as *"argument", "Whisper explaining", "criteria for explaining", "method of explaining ",* and *"purge or Whisper"*. Then, two researchers iteratively coded all the interview questions as related to each research question independently. Finally, three researchers discussed the themes and structures and mapped them out on the whiteboard.

3.3 Participant Demographics

Table 1 lists the main demographic characteristics of our participants. Most participants were male (71.5 %), followed by female (19%) and transgender (9.5%). The average age was 29, ranging from 18 to 45. The average moderation experience was two and a half years, ranging from one to five years. The number of channels they moderated ranged from one to eighty; however, most moderators moderated less than five channels (71%).The most active among participants had a channel list that contained 80 channels. Most are moderating gaming channels, and the viewership varies from hundreds to thousands.

4 RESULTS

Moderators applied a series of strategies to manage the content. We organized these strategies based on when they happen in relation to the bad act (Figure 2). The rectangular boxes represent a strategy. The straight lines represent a relation; the text on the straight line describes how they are related. The ovals represent an event. The diamonds represent when a decision needs to be made. The arrows represent a causal relation with the arrow pointing to the result, and the text on the arrow line represents the decision choice.

Following a time sequence, we presented the results from a proactive and reactive perspective with details such as why they used it, how they applied it, and in what situation they would use it to gain a comprehensive understanding of the moderation strategies in the live streaming community.

Table 1: Moderator Demographics and Moderation Activity

ID	No. of channels	Experience (yrs)	Age	Gender	Weekly (hrs)	No. of viewers	Channel type
P01	2	2-2.5	23	Male	21-84	10,000-60,000	Gaming
P02	1	2	-	Transgender	6	-	Board games
P03	6 or 7	5	31	Male	10	-	-
P04	80	4	24	Male	20	-	-
P05	30	3	21	Male	-	5-300	Gaming
P06	2	-	43	Male	Depends	few viewers	Gaming, products reviewing
P07	2	1	33	Female	20	70-130	Gaming and creative
P08	1	2	18	Male	60-70	10-100,000s	Gaming
P09	A couple	-	-	Male	35-42	2,000-30,000	-
P10	1	1.5	37	Female	3	-	Board games
P11	2	1	20	Male	21-28	5,000	Gaming
P12	1	1	21	Male	-	-	Gaming
P13	60	2.5	41	Male	21-28	500-600	Music and creative
P14	2 or 3	1	29	Male	12-16	-	-
P15	44	2	19	Male	2-3	700	Gaming
P16	20	2	40	Female	12	50-6,000	Gaming
P17	4	3-4	40	Male	4-12	200-7,000	Gaming
P18	3	4	-	Male	8-10	few viewers	Gaming
P19	5	5	27	Female	36-70	150-300	Gaming
P20	1	1	45	Transgender	16-24	100+	Gaming
P21	4	2	35	Male	30	100-500	Gaming

4.1 Proactive Strategies

Proactive strategies were ones that moderators engaged in before a viewer engages in a bad act and are represented in the top half of Figure 2, including 1) declaring presence, 2) rule echoing, 3) word blocking, and 4) setting a good example. In this section, we described the sequence and the interactions between the elements of the diagram, which are important to understand. We emphasized that moderators' work was complex but not arbitrary. The process began with monitoring without any intervention. If moderators felt that, possibly, the chatroom could potentially go wrong, they would intervene and say something to make moderators' presence in chat visible, which could deter the potential violators (declaring presence). At the same time, moderators could keep posting the rules and guidelines manually or through the bot in the chat to remind the newcomers (rule echoing). They would also activate the Twitch AutoMod to filter obvious toxic words (word blocking). If necessary, they interacted with viewers to set a good example so that other viewers could mirror their behaviors (setting a good example). Of importance, we found that setting a good example, rule echoing, and word blocking attempted to indicate norms while declaring presence, word blocking, and rule echoing attempted to deter potential violations.

4.1.1 Declaring presence. Declaring presence, as a method of deterring negativity before it happened, worked as an approach of gently reminding viewers that someone who had unique privileges to enforce the rules was monitoring the chat. Declaring their presence and showing viewers that they were active by only typing a word (moderators have a special sword symbol that supersedes their Twitch identifier) would curb and deter unwanted behaviors. P07 gave us an example:

> *If there was no active mod in there, people do try to push the lurk. They do say things that are inappropriate. Um, but when they see that there is even just one active mod, even if I just typed 'lol', they would see that there is a mod, that sort of cover for the trolls.*

This was a communicative strategy. Moderators showing their active status in the chat by simply replying or greeting viewers deterred potential norm violators. Unlike that of asynchronous communities, the "live" element of live streaming communities indicated that the moderator was watching on-site and that any following cross-border behaviors from violators could render punitive actions.

4.1.2 Rule echoing. The moderator had to actively and verbally inform viewers on a regular basis because even though rules were often displayed before someone had to type, they only automatically popped out once. Streamers usually had different rules for their channels. Some were obvious, such as no sexism, no harassment, no racism, and no profanity; others might involve prohibiting self-advertising and backseat gaming (which is spoiling the game for the streamer and other viewers). Therefore, posting rules in the channel was a way of proactively communicating these guidelines with the expectation that if the viewers saw them, they should follow them. P05 thought that, since the rules were posted, then they are clearly communicated, and expected the users to *"simply follow the rules"*. Yet, even if guidelines were posted on the channel, that did not mean that all users would read them. Newcomers often accidentally acted nonnormatively because they either did not know the rules

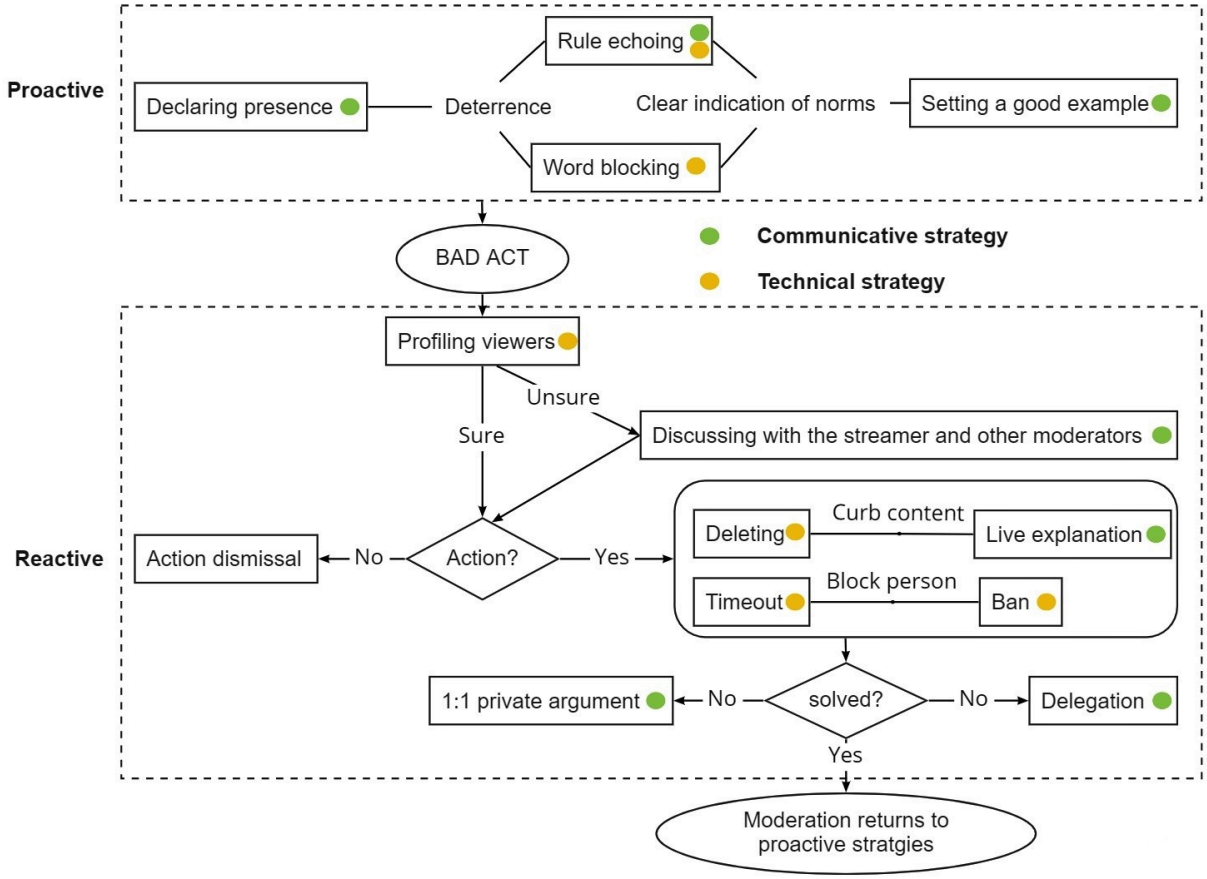

Figure 2: Moderation Strategies For Before and After a Bad Act Happens. Lines Indicate Relationships, Arrows Indicate Sequence.

or lacked experience [34]. Some moderators set up a bot that would be able to re-iterate the rules so that they would not have to type it out every time. For example, the command "!rules" would display the channel's guidelines. Using command or bot setting to post rules is both communicative and technical strategies visible to and for the public, clearly showing which behaviors are approved or disapproved.

4.1.3 Word blocking. Word blocking was achieved by the Twitch AutoMod that moderators could choose to activate to do some moderation tasks. AutoMod uses algorithms to hold inappropriate messages for moderators to review or prevent certain words from going into the chat. There are five levels (0 to 4) of moderation settings responding to moderation categories. Moderators could choose the moderation level and also update the terms under each level of the blocked terms list. A group of moderators reported that they liked the features of AutoMod because it could simply flag suspicious massages and reduce the workload to some extent.

If the messages were automatically filtered, only the moderator could see them until the messages were approved to the public chat so that other viewers would not be influenced. P18 expressed his appreciation for this feature:

> *By far my favorite feature of AutoMod is whenever people send a message it automatically doesn't go to the chat. It [AutoMod] pretends the message doesn't exist, it turns it into a none and done deal where no one saw it, no one is reacting, there's no drama- it's gone*

This was a technical strategy. The setting and application of AutoMod happened behind the scene, and the moderation process was invisible to the public. Applying moderation tools to block words is a common strategy that has been broadly discussed in online communities (e.g., [29, 51, 52]).

4.1.4 Setting a good example. Prior work has suggested that users want to fit in by doing what other community members tend to do (descriptive norms), and other community members' behaviors may be stronger indicators of acceptable ones than any explicit guidelines [34]. Similarly, we found that moderators reported being chatty, friendly, and *"answering questions"*(P19) as a way of keeping users positively engaged and hoping that viewers would imitate their behaviors. The moderator imperceptibly guided the viewers

to follow the rules through this method by showing what is the appropriate language and style in the chat, resonating with Seering et al.'s work [52]. P08 said, *"They kind of look up to me, kind of follow my lead."* Similarly, P05 said,*"In moderation, people look at you for what to do, how to act, and all that. So you have to always be talking, be chatting, be helpful to people, and especially off-stream you have to be that same personality."* According to P05, setting a good example was a communicative strategy involving more engagement and visibility in the public chatroom, showing a good personality as a community member and shaping the micro-community's value. Users imitating good behaviors supported a more enjoyable chat and reduced instances of banning.

While these were proactive strategies, we noted that these strategies could also be triggered by the reactive strategies discussed in the next section. For example, rule echoing could happen from a preventive perspective, but the moderators could also post rules after they ban or timeout the violators. In addition, word blocking could be updated after the moderators observed the lexical variations of toxic words.

4.2 Reactive Strategies

We identified nine reactive strategies as shown in the lower section of Figure 2. The novelty of our findings resided in the interaction and sequence of strategies. The process began when moderators observed bad actions that violated the rules. To avoid over-reactions and maintain the community, moderators would seek to understand viewers' behaviors by reviewing chat history or applying third-party tools to track viewers' chat messages (profiling viewers). If they understand the characteristics of these viewers, but they were unsure about the punishment they should give, they would ask other moderators or the streamer for help (discussing with the streamer and other moderators). If they were sure what they should do after profiling or after the discussion with other moderators and the streamer, they would decide to either dismiss the actions and ignore these messages (action dismissal) or take a series of actions to either curb the content (deleting or live explanation) or block the violators (timeout or ban). Sometimes, certain viewers were not satisfied with the punishment and would like to argue with the moderator privately (1:1 private argument). They could also keep harassing the stream with multiple accounts so that moderators had to delegate and ask the viewers to report the violator to the platform (delegation). Till then, the moderation process was completed and returned to proactive strategies.

In the following, we first introduced how moderators profiled viewers for decision making. Then we discussed other strategies with relevant quotes to explain each strategy such as why they would dismiss actions and ignore these messages, what the standards for blocking people and curb content were, and how they interacted with violators.

4.2.1 Profiling viewers. The purpose of profiling was to avert mistakenly blocking a person or curbing content because suppressing expression would hinder the growth of the community to some extent. Profiling could be very quick (several seconds) or sustain a very long time (varying from minutes to hours). It played a larger role in some situations than others. Moderators learned about viewers by either observing viewers' actions for several hours on a daily or weekly basis or reviewing the chat history and the specific viewer's history. Reviewing chat history was usually achieved through technical assistance difficult to obtain from the platform. Moderators had to use third-party tools that are allowed by Twitch to assist the profiling process. These third-party tools could provide more customization than AutoMod and allow moderators to track viewers' behaviors. P18 described a tool developed by his friend: *"His most useful tool by far is what he calls a log viewer, which pretty much lets me pull logs from anytime a user has talked in a channel as long as it's been logged."*

Especially when moderators had difficulty in deciding whether to take any action, checking the log would help them make better decisions. P5 explained, *"Whenever I see a new name in chat, I'll click them and see how long they've been on Twitch. if it's a day one account, I'm immediately skeptic and I watch them like a hawk. "* This information also helped moderators identify whether it was a repeated violator and decide whether it should be timed out or banned.

This was a technical strategy involving bot setting and operation to collect information. Prior work notes that moderators in user-governed communities apply various tools, including chat logs and post histories [52], but did not specify the purpose of these tools. We found that the account information and message history provided a background of the users that predicted their online behaviors. The information was helpful to the decision-making process when coming to moderation and improved moderation accuracy.

4.2.2 Discussing with the streamer and other moderators. Occasionally, a moderator did not know how to handle the situation and had to discuss the issue with the streamer or other moderators to finally *"mutually agree"* on how to deal with it, because they did not want to *"over-moderate"*. P01 explained,*"Like sometimes if we're not sure what to do [with] a person, we have a Skype chat and then we'll ask how we should deal with this person. Then we mutually agree on what to do with the person."* Similarly, P20 said,

> *If there are questionable situations, we'll have discussions among the moderators or with the streamers about what to do. In niche cases where we don't know about this, we have a discussion about it on Discord or in private message about what guidelines we want to have.*

Most of the time, they would directly discuss with other moderators first. Unless the situation was very serious, they would have to ask the streamers to make the final decision. P11 said,

> *I don't personally talk to the streamers. It's more kind of like a general knowledge thing if they tell you something like 'you don't have to ban this guy' or 'can you ban this guy?' 'can you time this guy out?', whatever. It's more of that kind of interaction. We don't personally have meetings with the streamers unless it's something super serious like a sponsorship or anything like that.*

Prior work has suggested that in user-governed online communities, moderators often apply an open discussion for changing rules in communities with less structured hierarchies, and the head moderators can arbitrarily make final decisions without asking for feedback in communities with a clear hierarchy [52]. We found that

in live-streaming communities, it was the streamer, not the head moderator or other moderators, making final decisions.

4.2.3 Action dismissal. After moderators had a basic understanding of the violators, they decided to ignore violations in some cases when they knew the viewers' persona, perceived viewers' intentions (to receive attention from others), or just decided to distance themselves from the situation.

Some viewers would always behave in a certain and expected pattern. In some situations, the moderator or the streamer had already classified these viewers' personas. With the streamer's approval, they decided to disregard these behaviors by doing nothing, even though those viewers violated the rules. P02 explained,

> *There's this guy. He likes to be toxic but then they're saying that's his personality online, like an online persona. It's just weird to me. It's just something I have to put up with.....Then I told [the streamer] about it and then [he] told me yeah that's just his personality. I said it's weird to me but okay.*

Some viewers broke the rules in order to get attention from others. Moderators elaborated that the best way to deal with these attention seekers was to ignore them and their inputs in the chat. *"Sometimes they're just looking for attention and sometimes you just ignore them; they just go away,"* said P19. The reason was that *"any further attention paid to them, it's just gonna feed them more. They're gonna continue trying to do it,"* said P17. Sometimes, the negative content caused heated discussion and increased the interaction in the chat. If the misbehavior was not very serious and the moderators thought it did not cross the line, they decided not to take action. P05 gave an example: *"Usually, if it's a really terrible troll I'll ignore them, then let them humiliate themselves and let chat have fun with it."*

We found that moderators used "let it go" as a strategy to distance themselves from the violator. P04 said, *"The easiest thing is if you have trolls trying to get through your skin you kind of let it go and laugh it off"*. Specifically, some moderators took short breaks to leave the screen and let these negative contents go with the chat flow instead of taking any further action. P07 shared her experience:*"I'm just going to go on a quick cup of tea. I'm having five minutes to myself and then went back."*

Action dismissal or non-response to violators is an atypical response to anti-normative behaviors. According to Figure 2, this is neither a technical nor communicative strategy, only involving cognitive processing. To a certain extent, high interactivity in the live chat results in the messages being transitory so that even though moderators did not take any action, the negative messages would disappear as more messages emerge. This strategy appropriately reduced information overload and emotional labor of moderators, but we did not know how the ignored content would affect other viewers. In order to minimize the negative impact of trolls, it has to be a widely followed norm of recognizing and ignoring them [34]. However, the challenge is that ignoring requires considerable self-control not to respond to offensive provocation, especially for new community members [26]. Thus, moderators may also need to educate viewers to identify and ignore these trolls, not just isolating themselves.

4.2.4 Live explanation. Recent work shows that Twitch users perceive educating as an effective strategy to get rid of toxicity [5]. Moderators explained the rules to viewers through live explanation and education. Unlike simply deleting with a warning, live explanation involved more engagement and offered help and suggestions to violators. The purpose of doing this was to build the community and curb the inappropriate content in the public chat without any punitive actions. Moderators often applied this strategy when they saw public argument among viewers or the chat topic became sensitive and was considered inappropriate for the public.

The public chat area is not a suitable place for arguments because it is mainly used for common topics that everyone can get involved with as well as interact with the streamer. An argument between two viewers could disturb the chat experience for other viewers as well. P08 said, *"If two people are arguing in the chat, I always [tell] them to take it to their DMs or Whispers or whatever to handle it there because the chat is not the place to do that."* P08's explanation is consistent with prior work that has indicated that moving conflicts to special locations where the normal rules of behavior do not apply will be met with less resistance from users [34]. The Whisper function of Twitch offers a private space for one-on-one interaction.

Though did not violate the rules, some topics were considered not appropriate in the chat because they were too personal or sensitive and could bring down the vibe and potentially cause negative impacts. Moderators dealt with those viewers by either providing resources they could utilize to help the viewers or politely asking them to leave, in an effort to protect the remaining viewers. P16 stated that she would remind these viewers to cease their actions:

> *There are some people who are negative because they're depressed. They come out like with their guns blazing and everything, and you tell them to knock it off, and they kind of back down pretty quickly. And you know, just speaking with them privately, you suggest that they get some help. I have phone numbers bookmarked for if people need someone to talk to, that sort of thing.*

The direct explanation between moderators and viewers could also rectify the misbehavior before it went beyond control and finally got the user banned. Moderators would gently remind the potential violators to remedy minor offenses. P20 said,

> *If they say something that they may not understand right. For example, sometimes people will walk in and will say something like, 'oh hey you're really pretty' and that's not an acceptable behavior so usually we will not ban them, we will say to them, 'hey that's objectifying and that's not an appropriate comment, it's not respectful to comment on the looks of a streamer so don't do that again'.*

P07 reported a similar tolerance: *"If they are less offensive and just being cheeky or maybe pushing a little bit, you send them a whisper and say, look, you know, calm down a little bit." "Usually the user will listen and apologize for it"*, P12 noted.

The educating and suggesting in both the public and private chat was a communicative strategy, either maintaining the chat atmosphere or rectifying lightweight violation. Prior work has discussed the black-box nature of the current moderation system and the lack of an educational system [43]. Live streaming communities

integrate the explicable and educational components into the moderation process. The synchronous nature of the live chat provides an opportunity for immediate feedback of the moderator's conduct to the viewer and also the viewer's performance to the moderator, making the education and explanation process possibly efficient. Our finding supplements Jhaver et al.'s work on Reddit that explanation of removal is under-utilized in moderation practices [30] and educating users with helpful feedback improves user attitude of fairness and intention to post in the future [28].

4.2.5 Deleting content, timeout, ban. These strategies were commonly applied as moderation activities. We found that, in "live" communities, deleting happened when the viewers did not read the rules of the chat and incidentally said something inappropriate. Even if moderators decided to remove these messages sometimes they did not ban the violator with an expectation that they would not perform the same behavior. P07 said, *"Those that just fail to understand what they're saying, it's either rude or something, we'll purge what they said."* Sometimes, deleting was followed by an explanation or warning, resonating Jiang's work of moderation in live voice communities [32].

Also, warning messages came in various forms of intensity. Some moderators used a gentle tone, reminding the viewers that such behaviors were not allowed, such as, *"Hey, we don't use that kind of language"*, said P12. Other moderators stated using severe sentences, cautioning users of the punishment awaiting them, should they proceed with their unacceptable actions. P03 said, *"You get that warning like 'Hey FYI, don't do this again otherwise you'll get ten-minute time out and then, you know, a third strike and you're banned'."*

A temporary ban, usually referred to as a "timeout", was a less severe solution for misbehavior compared with a permanent ban. Moderators reported having people in the chat who were mostly positive and respectful but might misbehave and cross the line. Temporarily banning the viewer who broke the rule sent a message to the viewer and the rest of the community that such behavior was not welcome. P05 stated, *"If you can tell someone has the intent of being a good community member, but they're a little overbearing, then that's a timeout."*

Several moderators deliberated that spamming emotes and text in the chat would get a timeout, which is different from Facebook that sends warning messages to the users and Twitter that investigates account activities, removes from search, or terminates the account [22]. P11 said, *"If someone is spamming the same message a couple of times, I will probably just time them off for ten minutes or so."*

A permanent ban meant that the users would never be allowed into the stream again. Not only was it a severe punishment for the user, but moderators also used this command sparingly because it affected the overall viewership. However, many moderators mentioned that they had *"zero tolerance"* towards obvious and severe issues such as racism and sexism and would ban these behaviors, similar to prior work [32, 52].

In addition, since live-streaming communities are streamer-centric, anything potentially harming the streamers and their benefits reserves severe punishment. Any personal attack toward streamers' appearances was also a permanent ban. Inappropriate comments such as the *"streamer's bad"* or the *"streamer's ugly"*, resulted in a permanent restriction on the viewer's ability to watch the stream (P08). P01 similarly reported, *"They might just like attack the players, whether physical appearance or lie how they play. Obviously, if it's physical appearance, then I have to purge them or ban them."*

One participant specifically mentioned that self-advertising of other streams deserves a permanent ban. In Twitch, many streams are similar in the content they provide, especially gaming streams. Thus, there is usually a lot of competition and a tendency to promote one's stream on other channels. P19 said, *"You actually do a permanent ban if they're advertising their stream in a chat. I don't have any type of tolerance or patience for that."* According to P19, the competition among different micro-communities escalates the moderation sanction. The content of self-advertising is not as severe as racism, sexism, or personal attack, but allowing it impairs the community, thus moderators have no *"tolerance or patience"*.

Sometimes moderators had different tolerance levels toward the same violation. For example, dealing with trolls in the chat was viewed differently by moderators. P21 said, *"You time someone out if they are troll. They will just leave because they don't want to wait ten minutes again and again."* But other moderators would permanently ban the same act. P06 said, *"But if someone is clearly just there to troll or just be a Jerk. Those people, there's nothing you can do with them, and there's no saving them. You just have to send them on their way."*

Generally, the deleting, timeout, and ban are technical strategies invisible to viewers and fitting the "graduated sanctions" [44], beginning with persuasion and light sanctions and proceeding to more forceful actions [34]. As parts of reactive strategies, the multi-level sanctions based on the severity of misbehaviors increases the legitimacy and thus the effectiveness of sanctions [34].

4.2.6 1:1 private argument. Viewers have the opportunity to argue with moderators through the private message; these conversations often happen during the stream. Sometimes viewers attempted to start arguments with moderators regarding the grey area between what was and was not allowed in a private chat. These arguments usually took place after a punitive action due to a user's misconduct in the chat. Viewers would argue that they should not be banned or timed-out through Whisper, and the moderators would argue the reason and deal with it on site. For example, P03 stated:

> *[The viewer is] being rude and being deliberately rude. Like the rules say don't be an XXX, and that's exactly what he was being... he kept bugging me, he's like 'well that doesn't really explain why you did what you did' and I said, 'Quite frankly, I'm here to do my job. I'm not here to be your friend.' I've said that before, and that's the ultimate thing.*

According to P03, the private chat allows the violator to express his opinion even after he was publicly banned. This process increased the perception of procedural justice, and the legitimacy is enhanced by providing users opportunities to argue their cases with the moderator [34]. The moderator was forced to perform in real-time in the private chat, which requires improvising. This was a communicative strategy, increasing the visibility of moderators in front of violators in the private chat. The nuanced difference between live explanation and private argument was that live explanation focused on the education and explanation in both public

and private chat while 1:1 private argument focused on the debate between moderators and violators in the private chat only.

4.2.7 Delegation. The moderators also encouraged other viewers to report violators because moderators could only process and deal with a limited amount of negative messages and problematic viewers even with the assistance of moderation tools. In certain situations, when the problematic viewer intentionally tried to disrupt the channel and created many accounts to harass the streamer or moderators, the moderator suffered from limited cognition and failed to address all issues in the chat. The information overload was difficult to handle in these situations. A smart approach to follow was to utilize the power of the crowds. Some moderators would ask viewers for support and do a "live crowdsourcing" to moderate chat comments. P13 said,

> *Maybe try to encourage viewers to go ahead and report this user, so hopefully, they get an IP ban. Those are only in really extreme cases when somebody won't go away, because Twitch is bad at that. If a viewer wants to create 50 accounts and harass someone privately, it's very hard to prevent that.*

This strategy was a communicative strategy seeking the public to engage, similar to moderation techniques encouraging users to flag suspicious content and report to the platform on Facebook [14] and relying on users as witnesses to collect evidence of rule-breakers in voice-based communities [32]. The reason behind this act was that volunteer moderators wished that the platform administrators (commercial moderators) could intervene since they might have more power to ban the IP of the violator.

5 DISCUSSION

This work mapped out the moderation strategies applied during the moderation process, contributing to the growing body of discussion about volunteer moderators and moderation in HCI and CSCW. We want to clarify that our main contribution is not the novelty of the strategies, but rather, it is the flow of how these strategies happen and the decision-making processes of moderators in the live context.

The interactivity of live streaming meant that moderators have to combine proactive and reactive strategies that engage both technical and communicative solutions, suggesting that moderators had to deal with harmful content in front of viewers on-site, explain and educate violators publicly or privately, and discuss with other moderators and the streamer behind the scene. These activities were accompanied by the challenge that because of the real-time nature, large volumes of content lead to information overload and only allow limited time for decision making and multi-task handling during the event. In the following section, we discuss how the unique affordances of live streaming increase the visibility of content moderation.

5.1 Interactivity Facilitates the Visible and Performative Activities of Moderation

Different from commercial content moderation that mostly happens behind the scenes [46] and lacks transparency [43], moderation relying heavily on volunteers increases the visible and performative activities. Among the 13 strategies in Figure 2, six involve technical, and seven involve communicative strategies. Technical strategies usually operate behind the screen and are less visible to viewers, while communicative strategies are mostly in the public chatroom visible to everyone or in private chat only visible to a specific violator. Only one strategy 'rule echoing' was found to fit both categories, where it is both a communicative and a technical strategy. Many communicative strategies applied at both proactive and reactive level can be achieved because live streaming provides an interactive and immersive experience for user engagement [24].

Moderators are usually the glorified viewers who are actively engaging in and influencing the channels [58]. During the streaming event, they still watch the stream as the viewers do, but with an eye on the chatroom. At the proactive level, the moderator sometimes needs to interact with viewers in the public chatroom by answering questions or joking around. This kind of performance happens in parallel to the performance of the streamer. In this sense, moderators are *the* viewers and *interacting with* other viewers. When they saw potential harmful actions, their roles would change to law enforcers who discretely dealt with the situation without disturbing the chat. They would either declare presence or post rules to deter these behaviors. Thus, the publicly visible activities involved different roles as moderators had to toggle between being the face of socialization/ community role model and justice enforcer. At the reactive level, the moderators have to explain and educate violators and delegate moderation tasks to viewers in the public chat or argue with violators in the private chat, indicating that the visibility of moderation increases moderators' vulnerability to negativity and violators [55]. How to balance moderation visibility and moderator protection should be further investigated.

Generally, volunteer moderators in the interactive context perform much visible communication in the public chat and private chat than commercial moderators do. The role (moderator, viewer) dynamic and visibility of volunteer moderators highlight the importance of affordances of live streaming when considering their roles and transparency and appear to be more prominent in the live streaming context in comparison to other social media platforms.

5.2 Synchronicity Enhances the Graduated Moderation and Amplifies the Violator's Voice

We echoed some moderation strategies broadly applied in online spaces such as content removal and banning the end-user [31, 54]. However, most of these common strategies are working separately. In most cases, one action is the end of moderation, such as content removal without rational explanation [28, 43] or directly banning the community [10].

According to the diagram in Figure 2, we systemically connected these moderation strategies and displayed them in a sequential flow to clearly show how moderation works in this new type of community. Kiesler et al. [34] applies the "graduated sanctions" concept in online community settings and suggested that the lowest level of sanctions is a private message explaining the violation, where sanctions escalate after repeated or more severe misbehavior. This concept can only partially explain connections of reactive

strategies but not proactive ones. Thus, "graduated moderation" seems to be more appropriate to include the proactive strategies in the workflow. The simultaneity and ephemerality of live streaming not only require instant attention and immediate moderation (e.g., one minute delay in moderation response could lead to a chaotic chat environment) but also make graduated moderation more effective than that on asynchronous communities because the moderators are always actively watching during the streaming event. The graduated moderation starting from proactive strategies instead of simply excluding violators shows the much effort moderators put to minimize the actions that could potentially alienate community members. Thus, graduated moderation increases the legitimacy and the effectiveness of moderation in the live context.

Moderation work in live-streaming communities empowers viewers to actively engage in the chatroom because the synchronicity brings everyone in the channel actively online all the time. The asynchronicity of most online communities limits the interaction of users and moderators and causes difficulty or delay for users to acquire feedback and guidance in time. Users lack the motivation to actively seek feedback unless moderators actively post explanations or contact the users. The delayed feedback discourages meaningful social engagement and relationship building. Prior work also points out that end-users develop their own folk theories configuring what is appropriate [15] because of the lack of explanation after content removal in online spaces [28]. In live-streaming communities, end-users can play larger roles than in asynchronous communities during the moderation process. For example, once a message was deleted, the viewer could ask the active moderators on-site for the reason or argue with the moderator that it was unfair. Thus, their voices can be heard by moderators in the dynamic interaction process and their valuable feedback may potentially contribute to the moderation process. Prior work has suggested that community influence on rule making increases compliance with the rules [34]. Therefore, community influence in live streaming plays a larger role on rule making than that in asynchronous communities, thus resulting in possibly higher compliance with the rules.

As new platforms emerge with novel technology, they may also take on property above currently unique to live streaming and consider how the moderation workflow works. For example, moderators in voice-based communities, such as Discord, secretly record voice for evidence and take extreme actions of excluding such as muting and banning [32]; instead of taking reactive strategies, moderators can combine some proactive strategies such as echoing the rules with declaring presence. The moderators can orally explain the rules or even have a recorded rule explanation to broadcast now and then in the voice channel. Though the diagram of moderation strategies is complex, it clearly shows the mental model of moderators. We can explicitly see where the decision making takes place and which strategy has been explored broadly or needs more attention.

5.3 Technological Implications

We propose that designers and developers should consider advanced technical tools to facilitate the profiling process. Current tools can only provide limited information about the viewers through the log function. Future tools should be able to provide more performance data of viewer's activity such as how long they have been online; how frequently these viewers communicate in the public chatroom and argue with moderators in the private chat; and according to these messages to tag the viewers' characteristics such as funny, talkative, elegant, well-behaved, toxic, and trolling, similar to the tagging mechanism on Twitter [27]. These data can help moderators increase the understanding of viewers and save time to make more accurate decisions during the moderation process.

An algorithm or system to identify the violators' type should be considered for moderators to make action dismissal decisions. We know that if the moderator knows the viewer's characteristics and intentions, they take no further action. For example, developers can design a classification system that can: (1) identify these problematic viewers based on text messages or chat history, (2) classify these viewers into specific categories such as attention seekers and a viewer saying bad words with good intention, and (3) annotate these messages and viewers and notify the moderators. This kind of system would reduce the monitoring effort and automatically catch violators when a large volume of messages pour into the chat, especially when moderators are handling a particular viewer and cannot keep an eye on the chat.

Communication is critical for effective moderation in live steaming communities, but the communication tools in the system were sub-par. We found that not all moderators would use the private messages function for discussion; they also used external tools such as Discord and Skype. Usually, a streaming channel has multiple moderators to ensure that at least one or two moderators are online when the streamer is. The problem, which is an opportunity for improving the design, is how the outcome of discussions between active moderators and the streamer can be documented so that other inactive moderators can be well-informed without wasting time checking the whole conservation history across different tools, which is simply an attempt to reinvent the wheel. It will be helpful if there is a system or feature that can automatically summarize the discussion in bullet points or highlights and save it as a document that can be shared with all moderators. Zhang and Cranshaw have developed a prototype system for Slack to automatically summarize chat conversation and share it with group members [62]. It is promising to bring such design to live streaming communities.

A documenting system would facilitate communication between not only moderators and streamers but also moderators and viewers. Explaining the rules through live interaction involves a lot of typing and interaction with viewers, which is time-consuming, and due to the limited cognitive abilities of the human brain, moderators might potentially overlook other negative content in the chat, causing a deterioration in the moderation job. If there is a bot or feature that can document these explanations in the system, and easily call out a specific explanation when necessary, we speculate that moderation efficiency would be highly improved by just simple 'click and send' instead of repeatedly typing. For example, we categorized 'rule echoing' as a communicative and a technical strategy. Since the content is already available in a written format, re-posting the relevant rule (as opposed to posting the entire rule list) when necessary, would help streamline the moderation process and increase the chances of viewers actually reading the automatic message.

5.4 Limitations

There are several limitations to this study. First, our participants are volunteers, not commercial moderators. In order to generalize the findings, further research can focus on commercial moderators in live streaming and compare the differences. Because the governance structure of each social media is different, we think it is inappropriate to claim that the user-moderated model in Twitch is similar to commercial moderation found in platforms like Facebook. Our findings may apply to other communities that have user-governance with simultaneity such as Discord, live-streaming communities, or live VR communities, but not all online communities. Also, even though our sample shows diverse moderation experience, we have more male than female and transgender. We are not sure if gender is something that influences moderation.

6 CONCLUSION

We identified the flow of decision-making that takes place during the moderation process. These practices of volunteer moderation bear similarities but also distinct differences compared with other user-governed communities. The interactivity and synchronicity of live streaming reveal the visible and performative work of volunteer moderation. This work reminds us to think about moderation from another perspective. Instead of considering moderation as blocking content or violators with the assistance of technical agencies, we may also want to take social dynamics into the moderation process and highlight the significance of communicative strategies performed by the human moderator at both the proactive and reactive level. The affordances of live streaming also allow graduated moderation and amplify violators' voices in the moderation process, showing moderators' great effort to increase legitimacy and maintain community members.

ACKNOWLEDGMENTS

This research was funded by by National Science Foundation (Award No. 1928627). Thanks to *ALL* the anonymous reviewers who has reviewed this manuscript for their insightful feedback. Thanks to the research assistants in SocialXLab at NJIT for data collection.

REFERENCES

[1] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *Proceedings of the 9th International Conference on Social Informatics*. 11. https://doi.org/10.1007/978-3-319-67256-4{_}32

[2] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (1 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[3] Simon Bründl and Hess Thomas. 2016. Why do users broadcast? Examining individual motives and social capital on social live streaming platforms. In *Proceedings of the 20th Pacific Asia Conference on Information Systems*. 332. https://aisel.aisnet.org/pacis2016/332/

[4] Jie Cai and Donghee Yvette Wohn. 2019. Categorizing Live Streaming Moderation Tools : An Analysis of Twitch. *International Journal of Interactive Communication Systems and Technologies* 9, 2 (2019), 36–50.

[5] Jie Cai and Donghee Yvete Wohn. 2019. What are effective strategies of handling harassment on twitch? Users' perspectives. In *Companion of the ACM Conference on Computer Supported Cooperative Work*. 166–170. https://doi.org/10.1145/3311957.3359478

[6] Jie Cai, Donghee Yvette Wohn, Ankit Mittal, and Dhanush Sureshbabu. 2018. Utilitarian and hedonic motivations for live streaming shopping. In *Proceedings of the 2018 ACM International Conference on Interactive Experiences for TV and Online Video - TVX '18*. New York, NY: ACM, 81–88. https://doi.org/10.1145/3210825.3210837

[7] Stevie Chancellor, Andrea Hu, and Munmun De Choudhury. 2018. Norms matter: Contrasting social support around behavior change in online weight loss communities. In *Conference on Human Factors in Computing Systems - Proceedings*. 1–14. https://doi.org/10.1145/3173574.3174240

[8] Stevie Chancellor, Jessica 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, CSCW*, Vol. 27. 1201–1213. https://doi.org/10.1145/2818048.2819963

[9] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 2. CSCW, 25. https://doi.org/10.1145/3274301

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

[11] Jonathan P. Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trajectories of blocked community members: redemption, recidivism and departure. In *Proceedings of the 2019 World Wide Web Conference*. 12. https://doi.org/10.1145/3308558.3313638

[12] Chia Chen Chen and Yi Chen Lin. 2018. What drives live-stream usage intention? The perspectives of flow, entertainment, social interaction, and endorsement. *Telematics and Informatics* 35, 1 (2018), 293–303. https://doi.org/10.1016/j.tele.2017.12.003

[13] Maxime Clément and Matthieu J. Guitton. 2015. Interacting with bots online: Users' reactions to actions of automated programs in Wikipedia. *Computers in Human Behavior* 50, 1 (2015), 66–75. https://doi.org/10.1016/j.chb.2015.03.078

[14] Kate Crawford and Tarleton Gillespie. 2016. What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media and Society* 18, 3 (2016), 410–428. https://doi.org/10.1177/1461444814543163

[15] Michael A. De Vito, Darren Gergle, and Jeremy Birnholtz. 2017. "Algorithms ruin everything": #RIPTwitter, folk theories, and resistance to algorithmic change in social media. In *Proceedings of 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 3163–3174. https://doi.org/10.1145/3025453.3025659

[16] Casey Fiesler, Jialun "Aaron" Jiang, Joshua McCann, Kyle Frye, and Jed R Brubaker. 2018. Reddit Rules! Characterizing an Ecosystem of Governance. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*. 72–81. https://doi.org/10.1016/j.tvjl.2007.05.023

[17] Mathilde B Friedländer. 2017. Streamer motives and user-generated content on social live-streaming services. *Journal of Information Science Theory and Practice* 55, 11 (2017), 65–84. https://doi.org/10.1633/JISTaP.2017.5.1.5

[18] Darren Geeter. 2019. Twitch created a business around watching video games - here's how Amazon has changed the service since buying it in 2014. https://www.cnbc.com/2019/02/26/history-of-twitch-gaming-livestreaming-and-youtube.html

[19] Ysabel Gerrard. 2018. Beyond the hashtag: Circumventing content moderation on social media. *New Media and Society* 20, 12 (2018), 4492–4511. https://doi.org/10.1177/1461444818776611

[20] Tarleton Gillespie. 2010. The politics of 'platforms'. *New Media and Society* 12, 3 (2010), 347–364. https://doi.org/10.1177/1461444809342738

[21] Tarleton Gillespie. 2017. Governance of and by platforms. In *SAGE Handbook of Social Media*, Jean Burgess, Thomas Poell, and Alice Marwick (Eds.).

[22] Tarleton Gillespie. 2018. *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, New Haven. 1–288 pages. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85051469782&partnerID=40&md5=8d850b5298b7e5dc1a1fcf4c427fe3da

[23] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data and Society* 7, 1 (2020), 1–15. https://doi.org/10.1177/2053951719897945

[24] Oliver L Haimson and John C Tang. 2017. What Makes Live Events Engaging on Facebook Live, Periscope, and Snapchat. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. 48–60. https://doi.org/10.1145/3025453.3025642

[25] William A Hamilton, Oliver Garretson, and Andruid Kerne. 2014. Streaming on twitch: fostering participatory communities of play within live mixed media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1315–1324. https://doi.org/10.1145/2556288.2557048

[26] Susan Herring, Kirk Job-Sluder, Rebecca Scheckler, and Sasha Barab. 2002. Searching for safety online: Managing "trolling" in a feminist forum. *Information Society* 18, 5 (2002), 371–384. https://doi.org/10.1080/01972240290108186

[27] Jane Im, Sonali Tandon, Eshwar Chandrasekharan, Taylor Denby, and Eric Gilbert. 2020. Synthesized Social Signals: Computationally-Derived Social Signals from Account Histories. In *Proceedings of the Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3313831.3376383

[28] 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. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3. CSCW, 33. https://doi.org/10.1145/3359294

[29] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. 2019. Human-Machine Collaboration for Content Regulation: The Case of Reddit Automoderator. *ACM Transition Human-Computer Interaction* (2019), 35. https://doi.org/10.1145/3338243

[30] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter?: User behavior after content removal explanations on reddit. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3. 27. https://doi.org/10.1145/3359252

[31] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online Harassment and Content Moderation: The Case of Blocklists. *ACM Transactions on Computer-Human Interaction* 25, 2 (2018), 1–33. https://doi.org/10.1145/3185593

[32] Jialun Jiang, Charles Kiene, Skyler Middler, Jed R Brubaker, and Casey Fiesler. 2019. Moderation challenges in voice-based online communities on discord. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3. 23. https://doi.org/10.1145/3359157

[33] Mladen Karan and Jaň Snajder. 2019. Preemptive Toxic Language Detection in Wikipedia Comments Using Thread-Level Context. In *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, Stroudsburg, PA, USA, 129–134. https://doi.org/10.18653/v1/W19-3514

[34] Sara Kiesler, Robert E. Kraut, Paul Resnick, and Aniket Kittur. 2012. Regulating Behavior in Online Communities. In *Building Successful Online Communities: Evidence-Based Social Design*, Robert E. Kraut and Paul Resnick (Eds.). The MIT Press, Chapter 4, 125–177. https://doi.org/10.7551/mitpress/8472.003.0005

[35] Kate Klonick. 2018. The new governors: The people, rules, and processes governing online speech. *Harvard Law Review* 131 (2018), 1598–1670.

[36] Jie Li, Xinning Gui, Yubo Kou, and Yukun Li. 2019. Live streaming as co-performance: Dynamics between center and periphery in theatrical engagement. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3. CSCW, 22. https://doi.org/10.1145/3359166

[37] Xiaocen Liu. 2016. Live streaming in China: boom market, business model and risk regulation. *Journal of Residuals Science & Technology* 13, 8 (2016), 1–7. https://doi.org/10.12783/issn.1544-8053/13/8/284

[38] Claudia Lo. 2018. *When All You Have is a Banhammer: The Social and Communicative Work of Volunteer Moderators*. Ph.D. Dissertation. Massachusetts Insitute of Technology. https://cmsw.mit.edu/banhammer-social-communicative-work-volunteer-moderators/

[39] Kiel Long, John Vines, Selina Sutton, Phillip Brooker, Tom Feltwell, Ben Kirman, Julie Barnett, and Shaun Lawson. 2017. "Could you define that in bot terms?": Requesting, creating and using bots on Reddit. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. 3488–3500. https://doi.org/10.1145/3025453.3025830

[40] Luck Marthinusen. 2017. Social media trends in 2018: Live streaming dominates the social media landscape. https://www.mo.agency/blog/social-media-trends-2018-streaming

[41] J. Nathan Matias. 2019. The Civic Labor of Volunteer Moderators Online. *Social Media + Society* 5, 2 (2019), 12. https://doi.org/10.1177/2056305119836778

[42] Aiden McGillicuddy, Jean Grégoire Bernard, and Jocelyn Cranefield. 2016. Controlling bad behavior in online communities: An examination of moderation work. In *2016 International Conference on Information Systems, ICIS 2016*. 11. https://slashdot.org/moderation.shtml

[43] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media and Society* 20, 11 (2018), 4366–4383. https://doi.org/10.1177/1461444818773059

[44] E. Ostrom. 1990. *Governing the commons: the evolution of institutions for collective action*. Cambridge University Press. 280 pages. https://doi.org/10.2307/3146384

[45] William Clyde Partin. 2019. Watch me pay: Twitch and the cultural economy of surveillance. *Surveillance and Society* 17, 1-2 (3 2019), 153–160. https://doi.org/10.24908/ss.v17i1/2.13021

[46] Sarah T Roberts. 2016. Commercial content moderation: Digital laborers' dirty work. In *The Intersectional Internet: Race, Sex, Class and Culture Online*, Safiya Umoja Noble and Brendesha Tynes (Eds.). NY: Peter Lang, New York, Chapter Commercial, 147–160. https://doi.org/10.1007/s13398-014-0173-7.2

[47] Sarah T. Roberts. 2017. Content moderation. https://doi.org/10.1007/3-540-35375-5

[48] Katrin Scheibe, Kaja J. Fietkiewicz, and Wolfgang G. Stock. 2016. Information behavior on social live streaming services. *Journal of Information Science Theory and Practice* 4, 2 (2016), 6–20. https://doi.org/10.1633/jistap.2016.4.2.1

[49] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. 2018. The Social Roles of Bots: Situating Bots in Discussions in Online Communities. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–29. https://doi.org/10.1145/3274426

[50] 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 - CSCW '17*. 111–125. https://doi.org/10.1145/2998181.2998277

[51] Joseph Seering, Michal Luria, Geoff Kaufman, and Jessica Hammer. 2019. Beyond dyadic interactions: Considering chatbots as community members. In *Proceedings of 2019 CHI Conference on Human Factors in Computing Systems - CHI ' 2019*. 13. https://doi.org/10.1145/3290605.3300680

[52] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media and Society* (2019), 1–28. https://doi.org/10.1177/ToBeAssigned

[53] Tim Squirrell. 2019. Platform dialectics: The relationships between volunteer moderators and end users on reddit. *New Media and Society* (2019). https://doi.org/10.1177/1461444819834317

[54] Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. In *Proceedings of the ACM on Human-Computer Interaction*, Vol. 3. CSCW, 21. https://doi.org/10.1145/3359265

[55] Lucy Suchman. 1995. Making Work Visible. *Communication of the ACM* 38, 9 (1995), 56–64. https://doi.org/10.4324/9781315648088

[56] T. Taylor. 2018. Twitch and the Work of Play. *American Journal of Play* 11, 1 (2018), 65.

[57] Andreas Veglis. 2014. Moderation techniques for social media content. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 8531. 137–148. https://doi.org/10.1007/978-3-319-07632-4{_}13

[58] Donghee Yvette Wohn. 2019. Volunteer moderators in Twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Proceedings of 2019 ACM Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3290605.3300390

[59] Donghee Yvette Wohn and Guo Freeman. 2020. Audience Management Practices of Live Streamers on Twitch. In *IMX 2020 - Proceedings of the 2020 ACM International Conference on Interactive Media Experiences*. Association for Computing Machinery, Inc, New York, NY, USA, 106–116. https://doi.org/10.1145/3391614.3393653

[60] Donghee Yvette Wohn, Guo Freeman, and Caitlin McLaughlin. 2018. Explaining viewers' emotional, instrumental, and financial support provision for live streamers. In *Proceedings of 2018 CHI Conference on Human Factors in Computing Systems*. ACM Press, New York, NY, 1–13. https://doi.org/10.1145/3173574.3174048

[61] Yu Chu Yeh. 2010. Analyzing online behaviors, roles, and learning communities via online discussions. *Educational Technology and Society* 13, 1 (2010), 140–151. http://www.jstor.org/stable/pdf/jeductechsoci.13.1.140.pdf

[62] Amy X. Zhang and Justin Cranshaw. 2018. Making sense of group chat through collaborative tagging and summarization. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 27. https://doi.org/10.1145/3274465