Pride and Professionalization in Volunteer Moderation: Lessons for Effective Platform–User Collaboration

Joseph Seering, Brianna Dym, Geoff Kaufman, and Michael Bernstein

Abstract. While most moderation actions on major social platforms are performed by either the platforms themselves or volunteer moderators, it is rare for platforms to collaborate directly with moderators to address problems. This paper examines how the group-chatting platform Discord coordinated with experienced volunteer moderators to respond to hate and harassment toward LGBTQ+ communities during Pride Month, June 2021, in what came to be known as the "Pride Mod" initiative. Representatives from Discord and volunteer moderators collaboratively identified and communicated with targeted communities, and volunteers temporarily joined servers that requested support to supplement those servers' existing volunteer moderation teams. Though LGBTQ+ communities were subject to a wave of targeted hate during Pride Month, the communities that received the requested volunteer support reported having a better capacity to handle the issues that arose. This paper reports the results of interviews with 11 moderators who participated in the initiative as well as the Discord employee who coordinated it. We show how this initiative was made possible by the way Discord has cultivated trust and built formal connections with its most active volunteers, and discuss the ethical implications of formal collaborations between for-profit platforms and volunteer users.

1 Introduction

Every June, LGBTQ+ communities across social media celebrate Pride Month, but with increased visibility comes an increase in targeted hate and harassment. For many volunteer-run communities on platforms like Reddit, Facebook groups, and Discord, the wave of hate is larger than group moderators are equipped to handle, even on platforms where the company takes an active role in removing homophobic and transphobic content. On June 1, 2021, the start of Pride Month, dozens of volunteer moderators from some of the largest communities on Discord, a group-chatting app popular among teenagers and young adults, teamed up with the company's Community Moderation Team to explore

10.54501/jots.v1i2.34

ways to support these communities. The volunteers, later nicknamed "Pride Mods," worked with Discord employees to identify servers that were likely to have difficulty handling the expected influx of hate and harassment and to offer an unusual form of support—for each server that requested help, Discord helped coordinate a set of highly trained volunteer moderators to temporarily join the server to bolster its existing moderation team and help with a variety of tasks. While having a few dozen volunteers join a handful of servers may seem less impactful than, for example, the company deploying a new moderation algorithm or hiring additional commercial content moderators, the intervention was very well received by all parties involved, including the servers' original moderation teams. Approximately 60 volunteers joined servers with a combined user population of more than one million, including several of the largest LGBTQ+ servers on Discord, allowing them to weather what would otherwise have been an overwhelming wave of hate.

Formal collaborations between platforms and users in the domain of moderation are rare. On platforms that rely in part on volunteer moderators, there is typically at least an implicit division of labor between the platform and the volunteers (J. Nathan Matias 2019b; Seering et al. 2019). For example, platforms frequently focus on broader issues and handle extreme or illegal content, while users deal with the moderation issues that arise within the communities they create (Seering 2020). Discord has pursued a much more collaborative approach to moderation by focusing on developing strong connections with users from communities with specific moderation needs. Discord has created what it calls a *Moderator Ecosystem*, a collection of moderation-focused Discord servers designed to give experienced volunteer moderators a direct line of contact with Discord employees, who collaborate with users to write educational resources for moderators and even to provide a formal program for semi-professional moderator training. The "Pride Mod" initiative described above grew from this collaborative approach. The idea originated from a conversation between volunteer moderators of large LGBTQ+ servers and the lead of Discord's Community Moderation Team, which is a part of the platform's Trust & Safety organizational structure. During the Pride Month initiative, this team lead directly recruited and coordinated "Pride Mod" volunteers from one of the Discord-run servers in the Moderation Ecosystem, connecting them with servers that had requested help.

Collaborative approaches to moderation have the potential to help address large-scale social problems that manifest very differently across contexts. For example, platforms have historically struggled to distinguish between obscenity and art or between violent, graphic images and political speech (see, e.g., Facebook's difficulty in determining how to moderate the famous *Terror of War* photograph (Roberts 2018; Seering 2020)). Though a very different problem, hate and harassment have proven similarly difficult for platforms to define in different contexts; language indicative of racism or homophobia in one context may have been co-opted to be humorous or even friendly when used within the targeted groups. While various platforms have attempted to find universal, context-agnostic responses to such problems, their attempts have consistently failed to achieve the desired results (Gillespie 2018; Oliver L. Haimson et al. 2021b). Community-based platforms that rely on volunteer moderators can allow users some leeway to define these lines within their own communities, partially addressing the challenge of moderating content differently in different contexts. However, this approach only works if volunteer moderators have the technical capacity and resources to enforce the rules they create. The "Pride Mod" initiative sought to tackle one instance of this problem: Discord as a company does not have the ability to moderate all homophobic, transphobic, and otherwise hateful language toward LGBTQ+ users with sufficient sensitivity to context, especially given the variety of ways in which group members have co-opted some of the slurs used against them. Similarly, volunteer moderators in large communities

may not always have the resources or time to moderate everything in their servers, particularly during especially busy periods (Jiang et al. 2019; Kiene, Monroy-Hernández, and Hill 2016; Kiene, Jiang, and Hill 2019). Volunteer moderators and Discord employees cooperated to at least partially solve both of these problems by creating a more effective, better-supported form of context-sensitive moderation in LGBTQ+ servers during Pride Month.

This paper chronicles the background, process, and outcomes of the "Pride Mod" initiative through interviews with the Discord Community Moderation Team employee who oversaw the efforts and 11 moderators who participated (including follow-up interviews with 10 of them). We analyze the initiative's effectiveness and consider the potential of this type of collaboration in a wider variety of contexts.

We explore four main questions:

1. How did the **context** of the Pride Mod initiative—LGBTQ+ focused servers—affect how it proceeded?
2. What were the **organizational challenges** associated with transplanting experienced moderators into new communities?
3. What **ethical questions** emerged from this use of volunteer labor by a for-profit platform?
4. Could this type of collaboration **generalize** to other contexts and platforms?

Broadly, we show that this initiative was successful in large part due to the ways in which Discord has cultivated trust with volunteer moderators and created its *Moderator Ecosystem*, which gathers and formalizes volunteers' expertise about moderation processes ranging from server setup to moderation team management to handling difficult situations. Discord was able to leverage these connections and the accumulated expertise to help coordinate support for communities in need of help during a difficult time. However, this form of platform–moderator collaboration would likely only succeed on platforms that have collaboratively developed a similar relationship with their moderators.

The Pride Mod initiative occurred at a time when major social media companies were increasingly focused on educating and training volunteer moderators.[1] As prior work has documented (Seering et al. 2019), formal training and onboarding materials for volunteer moderators are rare in the modern social media era; in some cases, communities create their own training documents, but platform support has typically been minimal. The recent increase in platform interest in training volunteer moderators signals a greater understanding of the value that these moderators bring, but it also raises questions about volunteer moderator labor. Volunteer moderators interviewed for this study often referred to their work as a second job; some regularly spend dozens of hours each week moderating their communities. The Pride Mod initiative was an especially clear example of this unusual labor arrangement, in which a company collaborated directly with a core set of highly trained volunteers to address a major moderation problem. This paper investigates the efficacy of this approach in the context of the four research questions to provide a starting point for more nuanced discussions of the ethics of this type of labor.

1. See, e.g., initiatives from Facebook, Reddit, and Discord.

2 Prior Work

This literature review discusses how previous studies have addressed three topics: (1) the social dynamics of community moderation, (2) the challenges of moderation in online LGBTQ+ spaces, and (3) interactions between users and platforms on issues of moderation. We address each of these in turn.

2.1 The Social Dynamics of Community Moderation

In public discourse, approaches to moderation are often grouped into two major categories: top-down (platform-driven) moderation and bottom-up (user-driven) moderation. The middle ground, in which users and platform employees both participate in moderation in different but often complementary ways, is common but frequently overlooked in academic literature (Seering 2020). For example, Reddit users are responsible for moderating individual communities; they are permitted to create rules, develop and use their own tools, and organize their moderation team as they see fit. Reddit has taken on the complementary responsibilities for handling certain types of extreme abuse and for making sure that moderators are maintaining a minimum standard of decency in their communities. Facebook groups exhibit a similar division of labor, though Facebook's moderation algorithms regularly supersede or contradict volunteer moderators' judgments. Even on YouTube, which is typically seen as an example of top-down moderation, users can moderate comments on their videos and livestreams or appoint other users to do so for them.

Recent work has documented the different roles that users and platforms play in content moderation, sometimes even noting the conflicts that take place between these groups (J. Nathan Matias 2019b, 2016). However, the academic literature has documented few cases of formal *collaboration* between users and platforms. Such cooperation may be unusual because relationships between platforms and volunteer moderators are generally rare or, in some cases, antagonistic. Such collaborations may also create expectations that platforms may wish to avoid. We argue in this paper that this form of collaboration, if properly designed, could help address significant shortcomings in current approaches moderation.

Online communities and their governance methods have been formally studied since the late 1970s (e.g., Hiltz and Turoff (1978)). Early research focused on the social potential and interpersonal behavioral challenges of the internet. There was a rapid increase in the volume of work studying the governance processes in online communities in the 1990s, including power dynamics (Reid 1999), problematic behaviors, (MacKinnon 1997; Donath 1999), moderation (Smith 1999), and community design (Morningstar and Farmer 1991). In the 2000s focus shifted toward productivity in online communities (e.g., Wikipedia (Kittur et al. 2007; Forte, Larco, and Bruckman 2009; Geiger and Ribes 2010) and open source software projects (O'Mahony and Ferraro 2007).

Researchers began focusing more and more on the "science" of moderation in the early 2010s, inspired in large part by Kraut and Resnick (2012)'s *Building Successful Online Communities*. Subsequent work in the social computing and computer-supported cooperative work fields has implicitly or explicitly addressed questions such as "What are the processes for moderation in online communities?" and "How can these processes be improved?" Seering et al. (2019), for example, mapped three such processes—"Being and becoming a moderator," "Moderation tasks, actions, and responses," and "Rules and community development." Prior studies have identified moderation challenges in specific types of spaces, including Asian-American and Pacific Islander communities on Reddit (Dosono and Semaan 2019) and special interest communities with high standards

for rigor in posted content (Gilbert 2020), and recent work has empirically tested different types of interventions (Seering, Kraut, and Dabbish 2017; Jhaver, Bruckman, and Gilbert 2019; J Nathan Matias 2019a). A separate strand of the literature challenges the implicit assumption in prior work that moderators were interchangeable generalists by identifying different archetypes for moderators who perform discrete roles within moderator teams (J. Nathan Matias 2019b; Wohn 2019; Seering, Kaufman, and Chancellor 2020).

This paper builds on these recent studies. For example, Seering et al. (2019)'s analysis of moderator onboarding processes could likely be extended to predict that the rapid onboarding of new moderators into unfamiliar communities would generally lead to poor outcomes. Similarly, we might draw from the conclusions in Dosono and Semaan (2019) and Gilbert (2020) to argue that moderation in identity-focused communities and spaces with fairly strict rules for permitted content and behaviors (both of which apply to the LGBTQ+ spaces we analyze here) would require a longer training period and/or a deep personal familiarity with the topic. Finally, we could conclude from past work on role division in moderation teams (Wohn 2019; Seering, Kaufman, and Chancellor 2020) that such teams in existing communities are likely to have balanced their labor across different roles, and an influx of new users without assigned roles would upset this balance. However, in the Pride Mod case, all of these predictions drawn from prior work, at least in the way we initially framed them, were incorrect. As we discuss below, our initial predictions underestimated how well moderators were able to bring in outside expertise and adapt it to these new contexts.

2.2 The Challenges of Moderating Online LGBTQ+ Spaces

Most of the studies mentioned above, with the exception of Dosono and Semaan (2019) and Gilbert (2020), drew broad, generalizable conclusions about volunteer moderation as a whole or across a specific platform rather than within specialized types of communities. While few studies have focused on volunteer moderation processes within LGBTQ+ communities on mainstream platforms like Reddit, Discord, or Facebook Groups, prior work has shown that different communities have different moderation needs and sometimes even different concepts of justice (Schoenebeck, Haimson, and Nakamura 2021). Guerrero Pico, Establés, and Ventura (2018) discussed moderation challenges when factions within a specific LGBTQ+ community become "toxic," but we are primarily interested in attacks from outsiders. Uttarapong, Cai, and Wohn (2021)'s work on the harassment experiences of women and LGBTQ+ live streamers on Twitch aligns more closely with the focus of this study, discussing the role of moderation tools and technologies as well as moderators' emotional and relational work.

Scheuerman, Branham, and Hamidi (2018)'s "Safe Spaces and Safe Places," though not written specifically about moderation, is perhaps the most comprehensive analysis of the harms that transgender people experience online. The authors situated harms across several axes—targeted vs incidental, directed toward individuals or entire communities, and sourced from outsiders or insiders—which align well with what interviewees shared with us in the current study. As discussed in Scheuerman, Branham, and Hamidi (2018), the moderators we interviewed described creating their servers to be a safe space for members of their communities. However, the desire to create a protective barrier, excluding outsiders who intended to do harm, often conflicted with the desire to recruit new well-intentioned users into the community, especially in times of increased visibility like Pride Month.

2.3 User–Platform Interactions in Moderation Contexts

Though few examples of collaborations between platforms and users on moderation issues have been documented in the research literature, prior work has identified numerous examples of conflicts. For example, Gerrard (2018) and Chancellor et al. (2016) identified users' strategies for finding and posting pro-eating disorder content on Instagram, circumventing the algorithms that platforms deploy to attempt to remove such content or prevent users from finding it. Other scholars have documented conflicts between users and platforms over rules about gender and self-presentation (Gerrard and Thornham 2020), including the years-long battle between Facebook and its users about the rules regulating breastfeeding photos Gillespie (2018). Similarly, Ruberg (2020) critiqued how Twitch defined sexual content, arguing that its rules regarding how women self-present (and how they are enforced) were based on inherent social biases against women in gaming spaces. Examples from Reddit illustrate a more direct form of collective action, in which moderators opposed to a particular action taken by the platform have organized "blackouts" with the (sometimes successful) intent to pressure the platform to make policy or technical changes (Centivany and Glushko 2016; J. Nathan Matias 2016).

Despite the relative dearth of studies specifically on volunteer moderation processes in LGBTQ+ online communities, much has been written about moderation-related *conflicts* between individual LGBTQ+ users and platforms. For example, Haimson and Hoffmann (2016) and MacAulay and Moldes (2016) investigated how Facebook's real names policy has harmed people with non-normative identities, including transgender and gender non-conforming users. Similarly, Edwards and Boellstorff (2020) documented conflicts between Tumblr and its users after it announced a ban on most adult content in late 2018, and Oliver L Haimson et al. (2021a) argued that these policy changes curtailed much of the freedom that had made the platform a "trans technology." A final thread of work on conflicts has analyzed the disproportionate impact that platform moderation and recommendation algorithms have on various populations of users. Simpson and Semaan (2021) argued based on interviews with LGBTQ+ TikTok users that, while the platform did in some cases reaffirm their identities by showing them relevant personalized recommendations, the algorithms tended to promote normative representations of those identities, e.g., "mainstream" lesbian content. Oliver L. Haimson et al. (2021b) found, based on user self-reports, that transgender users experienced disproportionately high levels of content removal, and Caplan and Gillespie (2020) referenced a number of cases in which LGBTQ+ users felt they had been targeted by YouTube's moderation algorithms because of their identities. Yet despite an extensive literature on conflicts, perhaps the most common form of interaction between users and platforms is a lack thereof. Seering et al. (2019) found that volunteer moderators in online communities frequently report that they have never communicated with any platform employees, and many feel the platform probably doesn't know that their community even exists.

3 Background

In May 2021, in preparation for an expected increase in hate speech and harassment on the platform during Pride Month 2021, Discord's Community Moderation Team began to connect with volunteer moderators, primarily from LGBTQ+ servers, to coordinate and discuss how best to provide support for servers that were likely to be targeted. On June 1, a representative of this team, which forms part of Discord's broader Trust & Safety organizational structure, sent a message to users in a Discord server that is accessible to owners of partnered servers, which are servers that have applied for formal recognition

and met certain standards for quality and activity.[2] Members of this server included moderators of LGBTQ+ focused servers as well as a variety of others.

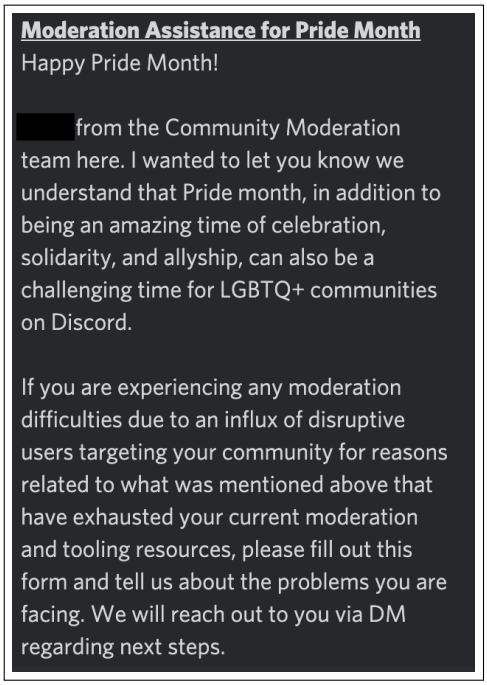

Figure 1: Message announcing opportunities to receive additional support. Post format reproduced from original text.

The message, shown in Figure 1, offered support for servers that expected to experience difficulties. It linked to a survey with questions about the name and type of server requesting help, what type of help would be needed, and when the additional volunteers should join.[3] Much of the coordination involved in matching volunteers to servers took place in a Discord server named the "Discord Moderator Discord" (DMD), which at the time of this study was run by Discord's Community Moderation Team to host and communicate with the most active and knowledgeable moderators on the platform. A similar announcement to the one in Figure 1 was made on the DMD, offering support to server moderators who felt they might need it, and another announcement was made in parallel soliciting volunteers who would *be* the support.

According to the Discord employee we interviewed, who we refer to as "Alex," preparations for the initiative began in April 2021 with discussions in the DMD between Discord staff and a group of server members, all of whom were experienced moderators—several on major LGBTQ+ focused servers.

> In 2020, Discord had promoted a few LGBTQ+ servers in [Server] Discovery, which increased activity in those servers and prompted support from the DMD at that time with [volunteer] moderators offering support. In 2021, we knew regardless of whether they appeared in discovery, LGBTQ+ servers would have increased activity regardless, so we wanted a more organized approach to democratize the support more, as well as provide a means for ensuring the moderators offering support could more easily coordinate (Alex, Discord Community Moderation Team).

2. Partner server requirements at the time of this study can be found here.
3. The full help request survey can be found at here.

Alex stated that it had been a priority for several years to recruit LGBTQ+ communities into the partner program discussed above and into the Moderator Ecosystem centered around the DMD, where moderators of LGBTQ+ servers would have a more direct line of communication to Discord Community Moderation Team employees. This aligns with what we found in our other interviews: most of the servers that requested support had at least one moderator in the DMD already, some of whom had helped plan the Pride Mod initiative.

It is important to note that most of the moderators who volunteered their support during the Pride Mod initiative were vastly more experienced than the average Discord moderator; many were working concurrently as moderators of some of the largest non-LGBTQ+ servers on Discord, and all were part of the Discord Moderator Ecosystem centered around the DMD. This ecosystem includes a variety of advanced moderation resources curated by Discord, including a formal training program run by its Community Moderation Team for volunteer moderators who want to moderate professional spaces, a mentorship program for younger moderators, and an "academy" hosted on the Discord Moderator Academy section of Discord's website. This academy features articles written by volunteer moderators (who are compensated for their time) on topics ranging from the details of bot management to high-level philosophical questions about moderation. Beginning in mid-2021, entry into this ecosystem of elite moderators required a top score on a moderation exam administered and graded by Discord Community Moderation Team employees, which featured questions on topics covered in the Discord Moderator Academy articles that were partly inspired by discussions within the DMD. Thus, the volunteer ecosystem members were highly experienced moderators who treated moderation as something approaching a formal profession.

A total of ten servers applied for support during Pride Month.[4] Three of the ten eventually decided they could manage without formal support, but the remaining seven all received volunteers; four were specifically LGBTQ+ focused, and three had a different focus but were receiving a large influx of hate due to Pride Month-specific events (e.g., hosting LGBTQ+ speakers, holding events focused on LGBTQ+ issues, and issuing pro-LGBTQ+ rights statements). These seven servers had a combined population of approximately 1.1 million members and included several of the largest LGBTQ+ focused servers on Discord, which, due to their size and visibility, were the most heavily targeted. More than 60 moderators from the DMD signed up to help and were distributed among the servers that requested support.

The initiative concluded at the end of June 2021. Some of the volunteers continued to moderate the communities they had joined, while others left to focus more on their other moderation work. Interviewees, including existing server moderators and new volunteers, almost universally agreed that the initiative had been successful overall and that the additional support provided by volunteers helped ease the burden of moderators who might otherwise have been overwhelmed.

4 Methods

We conducted interviews with eleven Discord moderators. Eleven interviews were conducted at the beginning of Pride Month and ten follow-up interviews were conducted with the same set of interviewees at the end of Pride Month. Interviewees came from six of the seven servers that requested and received help, either as pre-existing moderators for those servers or volunteers that joined. Six of the interviewees were pre-existing

4. Eight of the ten servers were primarily English speaking, one was Spanish speaking, and one was Portuguese speaking.

moderators, and five were new volunteers that joined servers that requested help. The first set of interviews focused on reasons for participating in the initiative (either in requesting or providing support), expectations for the process and its likely outcomes, and broader philosophical questions about moderator labor. The follow-up interviews focused on interviewees' opinions on the success of the initiative, the challenges involved, and additional questions about moderator labor to see if their opinions had changed after participating. Most interviews were conducted by Discord voice call, with an average interview length of 40 minutes, but some interviewees preferred to communicate via text or were not able to use audio.

We also conducted one structured interview via text with a Discord employee who was part of the Community Moderation Team that organized the initiative. In exchange for obtaining an on-the-record statement, the authors agreed to limit their questions to the Pride Mod initiative and relevant context. We believe this was a worthwhile compromise to make, but the resulting comments should be viewed with this in mind. The questions in this employee interview focused on the role Discord played in the initiative, the motivation and historical context for the initiative, the primary challenges encountered, and the place for this type of intervention in the future.

We report participants' demographics in aggregate because in some cases the combination of identity characteristics could clearly identify them to other members of their communities. We report their gender and sexual orientation in the language that participants used rather than attempting to standardize. Though some interviewees may have been transgender, and indeed some were moderators of trans-focused servers, we did not collect this information unless they specifically decided to identify as such.

Of the six pre-existing (PE) moderators, three identified as male (including one who specified that they are cis-male), two identified as non-binary, and one identified as female. Two moderators identified as heterosexual, one identified as gay, one as a lesbian, one as pan, and one as gray asexual and panromantic.

Of the five new volunteer (NV) moderators, four identified as male and one as female. Two identified as straight, one as gay, one as asexual, and one declined to share their sexual orientation.

All interviewees were aged 18–34; most were in their mid-20s. Nine were white, one was South Asian, and one was Hispanic. Seven were from the United States, three were from Western Europe, and one was from India.

The interview text was transcribed and the responses were grouped into categories based on the question being answered (or the topic if the interviewee was not explicitly answering a question). Open coding was then used to summarize common themes as per the guidelines for open coding in Creswell (2013, p. 86–89, 184–185). Text was separated into thematic chunks, each of which contained a single idea relevant to the research questions. Chunks varied in length from a few words to a full paragraph. The analysis was performed within pre-structured categories based on the research questions, but we inductively coded chunks into themes within each category. The results include summaries of the aggregated responses to direct questions as well as a discussion of emergent themes.

5 Results

In this section, we discuss our findings for each of the four research questions in turn, beginning with factors impacting how the intervention proceeded and moving to the ethical questions it raised and its broader applicability. We attribute quotes to the *type* of

interviewee (pre-existing (PE) or new volunteer (NV)) because the community of active moderators on LGBTQ+ Discord servers is small enough that it would be easy in some cases for members of that community to identify the interviewee, and we have chosen to prioritize their privacy.

5.1 Intervention Context: Moderation in LGBTQ+ Spaces

For the moderators who run them, LGBTQ+ servers are often incredibly meaningful and important spaces. One interviewee spoke about their offline experiences as a queer person, which involved being called slurs, being discriminated against at work, and even being the target of violence. They described how one of the servers they moderate was created after one of the founder's friends was killed because of their identity. The moderators of these servers are thus heavily invested in creating a truly safe space for people who may not have any other safe places to be themselves, a theme found in prior work on the creation of LGBTQ+ spaces on other platforms (Dym et al. 2019).

Accordingly, these moderators take their roles very seriously, in many such servers they are "on duty" in shifts that cover all hours of the day and night. They discussed the challenges of moderating LGBTQ+ servers at length in the interviews, identifying different types of common problematic behaviors they face both within the server and in direct messages they receive as moderators. The most common form of abuse was the use of homophobic and transphobic slurs, either by individuals or in "raids"—externally coordinated attacks against a server by a group of Discord users who intend to harass users en masse, cause chaos, and disrupt the normal function of the server. While most servers in this space have moderation bots configured to remove messages containing common slurs and their lexical variants, offending users are often creative in finding new ways to get around common rules. Slightly more sophisticated types of abuse include posting hateful imagery, memes, and sometimes videos, or using such images as profile photos. One moderator (PE) recalled a video that was posted showing a long list of companies simultaneously switching their logos back to non-pride colors as soon as Pride Month ended, sending a message to community members that everybody would stop caring about them as soon as July 1^{st} arrived. Though the deluge of hate was in many cases extreme, moderators of these spaces tended to discuss it with a fairly detached attitude.

> Most of the time, it's just another troll spamming slurs. You get used to it. Sometimes they hit those tough spots, but you generally just have to be ready for everything. (PE)

A major strategy for dealing with harassment and hate is the use of strict gating systems. On the majority of the LGBTQ+ servers moderated by interviewees, each user who wanted to join the server was required to go through a vetting process that involved answering a series of questions, often about their identity and how they found the server. In some cases they were required to provide links to their social media profiles so moderators could assess whether or not they were well intentioned. Along with each user's answers to these questions, some servers used custom-made bots to automatically gather data about the user, including when their account was created, whether they had verified the email address attached to the account, what invitation link they had used to join the server and who had created that link, and who else had joined using the same link. All of this information was then synthesized so that moderators could make a judgment about whether to allow the user to join, but in some cases moderators asked follow-up questions as well.

> Tons of people link us [social media] profiles where they're just blatantly homophobic, and it helps us keep out a lot of trolls, but we get like 600

> new users joining every day and with that many people joining, it's very time consuming. (PE)

Though this approach typically proved effective in preventing hate and harassment within the gated parts of the servers, it created an enormous separate workload for moderators who had to vet each user requesting to join.

A less visible form of attack reported by moderators in LGBTQ+ servers was harassment via direct messages. Most of the interviewees moderating these servers reported receiving messages full of slurs, death threats, and even images with extreme gore and child abuse.

> Of course we get that stuff. It just kind of goes without saying. Moderators understand what space we are in. (NV).

Moderators reported responding to this problem primarily in one of two ways: either turning off direct messages from users not on their friends list, or simply reporting the sender each time it happened. Moderators noted that the latter response could sometimes be difficult, as senders could simply delete their own message before it could be reported. Some moderators mentioned that Discord was in the process of internally testing ways to address this issue, but the lack of technical protections against this type of extreme personal attack remains a serious problem as of the time of writing.

5.2 Organizational Challenges: Transplanting Moderators into New Contexts

As described above, LGBTQ+ servers have unique moderation challenges, sophisticated moderation processes, and moderators who are often deeply emotionally invested in the server's well-being. Therefore, suddenly adding multiple moderators to the server who have no experience in these areas could be expected to cause serious problems. However, this was not the case in the context of the Pride Mod initiative. All Pride Mod volunteers reported that their onboarding experiences had gone at least fairly well, and all pre-existing moderators of servers that requested support reported that, aside from some confusion during the first few hours in a few cases, they were able to communicate well with the volunteers overall. In this section, we identify and discuss two possible reasons for this success.

First, the servers receiving support had fairly thorough onboarding materials for new moderators. These typically described the server's general moderation philosophy, the different roles in the server, how to use its moderation tools, and a list of example situations moderators might encounter and the procedure for handling them.

> Over the last year we reworked our training materials [...] We used to have a big Google Doc that was like a mod manual, but we threw that out and rebuilt it like a run book, like here are our procedures as a team and here's what we do in these situations. (PE)

Another moderator (PE), who was part of a moderation team with about 40 members in a very large gaming server, discussed what he felt had been a hectic onboarding process. In the afternoon of June 1^{st}, after realizing there was an urgent need for additional help, a moderator from this server was able to recruit 10 additional volunteer moderators through the Pride Mod initiative within half an hour. This moderator spent the next half hour explaining processes to the recruits, referring at times to existing documentation and ensuring that the recruits knew who to contact for various issues they might have.

> Our staff structure is a little bit different than most moderators are used to, so we had to explain that to them. We have a senior mod team with a

> manager, but we also have other people managing different teams like the events team. (PE)

While the new volunteers were tasked with banning users who used egregious slurs or hate speech, a moderator from this server wrote additional documentation to explain to the volunteers how to handle more nuanced, Pride-specific situations such as developing a process to deal with users from non-Western cultures for whom certain terms had different meanings. Though the organizational structure for moderators on this server was unusually complex, this level of organization made it easier for the server to absorb and rapidly integrate new moderators.

The second reason that the onboarding proceeded mostly smoothly was the level of prior experience and training that the volunteers had acquired as part of their prior moderator roles and their involvement in the Discord Moderator Ecosystem.

> This was easier compared to moderators who we have to train from scratch [...] since the moderators we requested for our [server] have the necessary experience and enthusiasm required for the job, it's much easier for them to just come in and help. (PE)

In the first round of interviews at the beginning of Pride Month, server moderators who requested support were asked whether they had concerns about non-LGBTQ+ volunteers stepping in to moderate an LGBTQ+-focused server. Most interviewees acknowledged this as a potential issue, but were optimistic that it would not be difficult to handle with good communication and documentation. In many servers, the new volunteers handled routine tasks such as vetting users and monitoring for hate and slurs not caught by the bots; while these tasks did require some subjective judgments, interviewees reported at the end of the month that they had been able to resolve issues that arose without significant problems by discussing them with pre-existing moderators. One server owner (PE) talked about a more social, interpersonal role for moderators, suggesting that they might be able to take on tasks like having private conversations with users having a particularly difficult time in their lives because of issues related to their identities. This moderator believed community members would be at least moderately receptive to having non-LGBTQ+ users participating in the community temporarily, in part because their presence constituted a strong signal of support for the community. In their follow-up interview at the end of the month, this PE reported that these types of positive personal interactions occasionally occurred between volunteers and community members did happen, though the frequency lower than expected.

The relative ease of onboarding during this initiative raises questions about the interchangeability of moderators. For example, is moderation work perhaps not as community specific as prior research has led us to believe? The response to this question likely has two parts. First, the moderators involved in this initiative were among the best trained on the platform, within the top few hundred among hundreds of thousands. It is very unlikely that an average moderator could adapt to a new environment so quickly. The second part of the response is that while the volunteers did successfully integrate into the new context, they were primarily placed in roles in which they dealt with content (like aggressive slurs and harassment) that is egregious enough to be handled similarly across most well-run servers and for which specialized knowledge was not needed to determine how to intervene. These volunteers did not typically participate in the broader governance of the servers they joined, nor did they make new policies or decisions about the future of the server. Thus, while it is plausible that there is a core set of skills that are transferable across different types of communities, it remains likely that certain roles are best filled by moderators with community-specific experience.

5.3 Ethical Questions Associated with the Use of Volunteer Labor

Though the Pride Mod initiative was widely regarded by all interviewees as a success, at its core it involved using unpaid volunteer labor to solve a moderation challenge on a for-profit platform. The initiative's success must therefore be judged in the context of questions about the ethics of a company's paid employees working alongside unpaid volunteers on one of its most complex and challenging problems.

> In all this, we learned that we have the ability to connect willing moderators with communities in need of assistance, and *we should leverage not only the tools we offer* for moderation in the platform *and the work our internal teams do* more generally when there are notable influxes of users to any cluster of communities, *but* [also] *our community at large* [emphasis added].
>
> We also learned that there are a lot of people out there who care about platform safety in the sense that they are willing to help out their favorite communities, and communities they may be learning more about for the first time, all while bringing their skills and expertise into what it takes to build safe and thriving communities. (Alex, Discord Community Moderation Team.)

This strategy could be criticized as inherently unethical: one could argue that instead of spending its own resources, Discord offloaded its responsibilities onto volunteers who contributed thousands of hours of unpaid labor (Brown and Hennis 2019). Though this argument is worth taking seriously, every volunteer moderator interviewed for this study maintained that the initiative was not exploitative at its core for two reasons. First, it would be financially impossible for Discord to hire moderators to moderate every server. Second, even if Discord could, that would not be a desirable outcome. The moderators argued that the situation could not be labeled as exploitative or not on its own, but should be considered in comparison to possible alternatives. Given the choice between being allowed to moderate their own communities (albeit through significant unpaid labor) and *not* being allowed to moderate their own communities, interviewees strongly preferred the former.

One interviewee added a caveat to this position, arguing that a volunteer-based moderation system can only be ethical so long as the platform listens to the needs of moderators and spends significant resources developing tools and processes to support them.

> I don't really think that it's Discord's responsibility to pay mods, but I think Discord can do more [to support us]. [For example,] Discord doesn't really offer any type of psychological support or or mental support, but I regularly have to look at and delete very inappropriate [content] like videos of suicides, for example, and of child [sexual abuse]. (PE)

Ultimately, it is important to remember that Discord is not a platform moderated only by volunteers; there is a division of labor between volunteers and the platform in which each performs a set of distinct but complementary tasks. Rather than asking "Is volunteer moderation ethical?", we argue that it is more productive to ask questions like "Is the division of labor fair?" and "Are volunteer moderators given the support they need to safely and effectively manage their communities?" Treating volunteer-reliant models as part of a spectrum of possible models rather than as a monolith can lead to more productive discussions about how they might be improved in the future.

5.4 Generalizability of the Pride Mod approach

The Pride Mod approach to handling harassment and hate at scale proved effective in this case. Based on the interviewees' responses, it could also be effective in other similar

situations on Discord. Alex, the Community Moderation Team employee responsible for overseeing the initiative, confirmed that there are plans to expand beyond the Pride Mod case to create a more formal, ongoing program that operates in much the same way:

> As our safety programs grow and we certify more moderators within programs, we are confident there will be a broader base of moderators willing to participate in this program. We also will be able to reach out to more communities—in and outside of our programs—more in advance, so we anticipate the program will expand. (Alex, Discord Community Moderation Team)

In their interview, Alex strongly emphasized that they thought "Finding communities wanting this assistance!" was the biggest challenge associated with making an initiative like this successful. Discord servers do not have a single central directory, and many servers are intentionally kept private without a public footprint. As such, the company cannot feasibly contact moderators of every server that might benefit from support during such an initiative. In this initiative, the servers that received an offer of support were primarily of sufficient size and activity to participate in formal Discord programs (and most were English-language servers), though some additional servers were contacted through less formal means. Though the most active servers are by definition the most likely to need support, since attackers are more likely to discover and target them, we suggest that future initiatives should be assessed partly according to their ability to reach as many communities in need of assistance as possible.

Alex and many of the moderator interviewees were broadly optimistic about the potential future for this type of program on Discord. However, evidence gathered in our interviews suggests that implementing such a program on another platform might be challenging for two reasons: the need for pre-existing server moderators to trust the volunteers, and the need for both to trust the host platform. This program was effective in large part due to Discord's focus on professionalizing volunteer moderation through its Moderator Ecosystem. The volunteer Pride Mods were very experienced but, perhaps more importantly, they were part of an established community with strong reputation signals. Many of them already knew each other, and those who did not had a wealth of information they could draw on to evaluate each other. Server owners were much more comfortable trusting these moderators to join their communities than they would have been with users they had no connection to because they could be fairly confident of the Pride Mods' good intentions and competence. Most social platforms do not have communities like the Discord Moderator Ecosystem, which are far from trivial to create; Discord has spent several years and most likely tens of thousands of hours curating and growing this community and the associated resources. Though Alex was optimistic when asked about the potential for this approach to work on other platforms, they acknowledged the importance of having such a community:

> Discord was able to leverage networks of users within our education initiatives and communities, so a similar means of connecting with users on the platform would be instrumental in organizing a comparable effort. (Alex, Discord Community Moderation Team)

Choosing to participate in such an initiative requires a non-trivial amount of trust in the host platform, particularly for moderators of LGBTQ+ communities who might have concerns about granting authority to moderators who had been recruited and organized by the platform. Yet when the Discord Pride Mod initiative was announced, owners of several major servers were willing to trust the process enough to request support. This trust was built over the course of several years of mostly positive interactions between experienced moderators and Discord employees; most interviewees could recall at least

one instance in which they had discussed a problem with Discord staff and had felt their concerns were heard. For example, several interviewees recalled a situation in which an issue with a newly announced feature could have led LGBTQ+ users to accidentally out themselves if they used certain server-specific emojis in the wrong context, but interviewees reported that Discord staff were responsive to this concern when made aware of it and the feature was adjusted. Some interviewees even mentioned cases in which Discord employees joined servers to help moderate in special cases when urgent support was requested.

> So I do think really, if it did absolutely come down to it, the [Discord] staff that are working on outside stuff or working on support, trust and safety or development, they would take the time to help. (PE)

However, users on other platforms might have less trust in employees who reached out to offer support. Group moderators on Facebook, for example, might be more skeptical about whether Facebook had ulterior motives if it announced an initiative like this one.

Broadly, though there is significant potential for cooperation between users and platforms to help address serious moderation problems that neither party has been able to solve alone, such collaborations require a curated social infrastructure and a significant amount of trust in order to succeed.

6 Conclusion

This article documents a rare example of a successful collaboration between users and a platform to address a serious moderation issue. We have discussed how, while LGBTQ+ communities face unique moderation challenges due to the volume and variety of hate they receive, especially during Pride Month, an experienced set of volunteers was able to successfully integrate into their moderation teams to provide temporary support. We have also argued that, while there are reasonable questions about the ethics of a for-profit platform relying on volunteer labor, future debates over the ethics of volunteer-reliant models for moderation would benefit from comparing different forms this model can take rather than treating it as a homogeneous approach to moderation. Finally, we have argued that this approach could work on other platforms, but that there are significant implementation challenges that would require a long-term investment of resources to overcome.

The most effective solutions to the most pressing moderation issues across the social web are likely to require participation from users as well as platforms; joint efforts could accomplish more than either actor could on its own. Though there are many ways in which platforms and users might collaborate in the future, it is important to consider what models are the most promising while remaining ethically sound.

References

Brown, James J, and Gregory Hennis. 2019. "Hateware and the outsourcing of responsibility." In *Digital Ethics,* 17–32. Routledge.

Caplan, Robyn, and Tarleton Gillespie. 2020. "Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy." *Social Media + Society* 6 (2): 2056305120936636. https://doi.org/10.1177/2056305120936636. https://doi.org/10.1177/2056305120936636.

Centivany, Alissa, and Bobby Glushko. 2016. ""Popcorn Tastes Good": Participatory Policymaking and Reddit's," 1126–37. CHI '16. San Jose, California, USA: Association for Computing Machinery. https://doi.org/10.1145/2858036.2858516. https://doi.org/10.1145/2858036.2858516.

Chancellor, Stevie, 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 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing,* 1201–13. CSCW '16. San Francisco, California, USA: Association for Computing Machinery. https://doi.org/10.1145/2818048.2819963. https://doi.org/10.1145/2818048.2819963.

Creswell, John W. 2013. *Qualitative Inquiry and Research Design: Choosing Among Five Traditions.* Thousand Oaks, CA: Sage.

Donath, Judith. 1999. "Identity and Deception in the Virtual Community." In *Communities in Cyberspace,* 1st, edited by Marc A Smith and Peter Kollock, 27–58. London, UK: Routledge. https://doi.org/10.1519/JSC.0b013e3181e4f7a9.

Dosono, Bryan, and Bryan Semaan. 2019. "Moderation Practices as Emotional Labor in Sustaining Online Communities: The Case of AAPI Identity Work on Reddit." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery. https://doi.org/10.1145/3290605.3300372. https://doi.org/10.1145/3290605.3300372.

Dym, Brianna, Jed R. Brubaker, Casey Fiesler, and Bryan Semaan. 2019. ""Coming Out Okay": Community Narratives for LGBTQ Identity Recovery Work." *Proc. ACM Hum.-Comput. Interact.* (New York, NY, USA) 3, no. CSCW (November). https://doi.org/10.1145/3359256. https://doi.org/10.1145/3359256.

Edwards, Emory James, and Tom Boellstorff. 2020. "Migration, non-use, and the 'Tumblrpocalypse': Towards a unified theory of digital exodus." *Media, Culture & Society* 0 (0): 0163443720968461. https://doi.org/10.1177/0163443720968461. https://doi.org/10.1177/0163443720968461.

Forte, Andrea, Vanesa Larco, and Amy Bruckman. 2009. "Decentralization in Wikipedia Governance." *Journal of Management Information Systems* 26 (1): 49–72. https://doi.org/10.2753/MIS0742-1222260103. https://doi.org/10.2753/MIS0742-1222260103.

Geiger, R. Stuart, and David Ribes. 2010. "The Work of Sustaining Order in Wikipedia: The Banning of a Vandal." In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work,* 117–26. CSCW '10. Savannah, Georgia, USA: ACM. https://doi.org/10.1145/1718918.1718941. http://doi.acm.org/10.1145/1718918.1718941.

Gerrard, Ysabel. 2018. "Beyond the hashtag: Circumventing content moderation on social media." *New Media & Society* 20 (12): 4492–511. https://doi.org/10.1177/1461444818776611. https://doi.org/10.1177/1461444818776611.

Gerrard, Ysabel, and Helen Thornham. 2020. "Content moderation: Social media's sexist assemblages." *New Media & Society* 22 (7): 1266–86. https://doi.org/10.1177/1461444820912540. https://doi.org/10.1177/1461444820912540.

Gilbert, Sarah A. 2020. ""I Run the World's Largest Historical Outreach Project and It's on a Cesspool of a Website." Moderating a Public Scholarship Site on Reddit: A Case Study of r/AskHistorians." *Proc. ACM Hum.-Comput. Interact.* (New York, NY, USA) 4, no. CSCW1 (May). https://doi.org/10.1145/3392822. https://doi.org/10.1145/3392822.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media.* New Haven, CT, USA: Yale University Press.

Guerrero Pico, Mar, María-José Establés, and Rafael Ventura. 2018. "Killing off Lexa: 'Dead lesbian syndrome' and intra-fandom management of toxic fan practices in an online queer community." *Participations* 15:311–33.

Haimson, Oliver L, Avery Dame-Griff, Elias Capello, and Zahari Richter. 2021a. "Tumblr was a trans technology: the meaning, importance, history, and future of trans technologies." *Feminist Media Studies* 21 (3): 345–61. https://doi.org/10.1080/14680777.2019.1678505. https://doi.org/10.1080/14680777.2019.1678505.

Haimson, Oliver L., Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021b. "Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas." *Proc. ACM Hum.-Comput. Interact.* (New York, NY, USA) 5, no. CSCW2 (October). https://doi.org/10.1145/3479610. https://doi.org/10.1145/3479610.

Haimson, Oliver L., and Anna Lauren Hoffmann. 2016. "Constructing and enforcing "authentic" identity online: Facebook, real names, and non-normative identities." *First Monday* 21, no. 6 (June). https://doi.org/10.5210/fm.v21i6.6791.

Hiltz, Starr Roxanne, and Murray Turoff. 1978. *The Network Nation: Human Communication via Computer.* Boston, MA: Addison-Wesley Publishing Company, Inc.

Jhaver, Shagun, Amy Bruckman, and Eric Gilbert. 2019. "Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit." *Proc. ACM Hum.-Comput. Interact.* (New York, NY, USA) 3, no. CSCW (November). https://doi.org/10.1145/3359252. https://doi.org/10.1145/3359252.

Jiang, Jialun Aaron, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2019. "Moderation Challenges in Voice-based Online Communities on Discord." *Proc. ACM Hum.-Comput. Interact.* (New York, NY, USA) 3, no. CSCW (November): 55:1–55:23. https://doi.org/10.1145/3359157. http://doi.acm.org/10.1145/3359157.

Kiene, Charles, Jialun Aaron Jiang, and Benjamin Mako Hill. 2019. "Technological Frames and User Innovation: Exploring Technological Change in Community Moderation Teams." *Proc. ACM Hum.-Comput. Interact.* (New York, NY, USA) 3, no. CSCW (November): 44:1–44:23. https://doi.org/10.1145/3359146. http://doi.acm.org/10.1145/3359146.

Kiene, Charles, Andrés Monroy-Hernández, and Benjamin Mako Hill. 2016. "Surviving an "Eternal September": How an Online Community Managed a Surge of Newcomers." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems,* 1152–56. CHI '16. San Jose, California, USA: ACM. https://doi.org/10.1145/2858036.2858356. http://doi.acm.org/10.1145/2858036.2858356.

Kittur, Aniket, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. 2007. "He Says, She Says: Conflict and Coordination in Wikipedia." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 453–62. CHI '07. San Jose, California, USA: Association for Computing Machinery. https://doi.org/10.1145/1240624.1240698. https://doi.org/10.1145/1240624.1240698.

Kraut, Robert, and Paul Resnick, eds. 2012. *Building Successful Online Communities: Evidence-based Social Design.* Cambridge, MA, USA: MIT Press.

MacAulay, Maggie, and Marcos Daniel Moldes. 2016. "Queen don't compute: reading and casting shade on Facebook's real names policy." *Critical Studies in Media Communication* 33 (1): 6–22. https://doi.org/10.1080/15295036.2015.1129430. https://doi.org/10.1080/15295036.2015.1129430.

MacKinnon, Richard. 1997. "Virtual Rape." *Journal of Computer-Mediated Communication* 2 (4): 1–2. https://doi.org/10.1111/j.1083-6101.1997.tb00200.x.

Matias, J Nathan. 2019a. "Preventing harassment and increasing group participation through social norms in 2,190 online science discussions." *Proceedings of the National Academy of Sciences* 116 (20): 9785–89. https://doi.org/10.1073/pnas.1813486116. https://www.pnas.org/content/116/20/9785.

———. 2016. "Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout," 1138–51. CHI '16. San Jose, California, USA: Association for Computing Machinery. https://doi.org/10.1145/2858036.2858391. https://doi.org/10.1145/2858036.2858391.

———. 2019b. "The Civic Labor of Volunteer Moderators Online." *Social Media + Society* 5 (2). https://doi.org/10.1177/2056305119836778.

Morningstar, Chip, and F Randall Farmer. 1991. "The lessons of Lucasfilm's habitat." In *Cyberspace,* 273–302. MIT Press.

O'Mahony, Siobhán, and Fabrizio Ferraro. 2007. "The Emergence of Governance in an Open Source Community." *Academy of Management Journal* 50 (5): 1079–106. https://doi.org/10.5465/amj.2007.27169153. https://doi.org/10.5465/amj.2007.27169153.

Reid, Elizabeth. 1999. "Hierarchy and Power: Social Control in Cyberspace." In *Communities in Cyberspace,* 1st, edited by Marc A. Smith and P. Kollock, 107–34. New York, NY, USA: Routledge.

Roberts, Sarah. 2018. "Digital detritus: 'Error' and the logic of opacity in social media content moderation." *First Monday* 23 (3). https://doi.org/10.5210/fm.v23i3.8283. https://journals.uic.edu/ojs/index.php/fm/article/view/8283.

Ruberg, Bonnie. 2020. ""Obscene, pornographic, or otherwise objectionable": Biased definitions of sexual content in video game live streaming." *New Media & Society,* https://doi.org/10.1177/1461444820920759. https://doi.org/10.1177/1461444820920759.

Scheuerman, Morgan Klaus, Stacy M. Branham, and Foad Hamidi. 2018. "Safe Spaces and Safe Places: Unpacking Technology-Mediated Experiences of Safety and Harm with Transgender People." *Proc. ACM Hum.-Comput. Interact.* (New York, NY, USA) 2, no. CSCW (November). https://doi.org/10.1145/3274424. https://doi.org/10.1145/3274424.

Schoenebeck, Sarita, Oliver L Haimson, and Lisa Nakamura. 2021. "Drawing from justice theories to support targets of online harassment." *New Media & Society* 23 (5). https://doi.org/10.1177/1461444820913122. https://doi.org/10.1177/1461444820913122.

Seering, Joseph. 2020. "Reconsidering Self-Moderation: The Role of Research in Supporting Community-Based Models for Online Content Moderation." *Proc. ACM Hum.-Comput. Interact.* (New York, NY, USA) 4, no. CSCW2 (October). https://doi.org/10.1145/3415178. https://doi.org/10.1145/3415178.

Seering, Joseph, Geoff Kaufman, and Stevie Chancellor. 2020. "Metaphors in moderation." *New Media & Society,* https://doi.org/10.1177/1461444820964968. https://doi.org/10.1177/1461444820964968.

Seering, Joseph, 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,* 111–25. CSCW '17. Portland, Oregon, USA: ACM. https://doi.org/10.1145/2998181.2998277. http://doi.acm.org/10.1145/2998181.2998277.

Seering, Joseph, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. "Moderator engagement and community development in the age of algorithms." *New Media & Society* 21 (7): 1417–43. https://doi.org/10.1177/1461444818821316.

Simpson, Ellen, and Bryan Semaan. 2021. "For You, or For"You"? Everyday LGBTQ+ Encounters with TikTok." *Proc. ACM Hum.-Comput. Interact.* (New York, NY, USA) 4, no. CSCW3 (January). https://doi.org/10.1145/3432951. https://doi.org/10.1145/3432951.

Smith, Anna DuVal. 1999. "Problems of Conflict Management in Virtual Communities." In *Communities in Cyberspace,* 1st, edited by Marc A Smith and P Kollock, 135–66. New York, NY, USA: Routledge.

Uttarapong, Jirassaya, 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,* 7–19. IMX '21. Virtual Event, USA: Association for Computing Machinery. https://doi.org/10.1145/3452918.3458794. https://doi.org/10.1145/3452918.3458794.

Wohn, Donghee Yvette. 2019. "Volunteer Moderators in Twitch Micro Communities: How They Get Involved, the Roles They Play, and the Emotional Labor They Experience." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,* 160:1–160:13. CHI '19. Glasgow, Scotland Uk: ACM. https://doi.org/10.1145/3290605.3300390. http://doi.acm.org/10.1145/3290605.3300390.

Authors

Joseph Seering is a postdoctoral scholar at Stanford University.

Brianna Dym is a PhD candidate at the University of Colorado, Boulder.

Geoff Kaufman is the Robert E. Kraut Associate Professor of Human–Computer Interaction at Carnegie Mellon University.

Michael Bernstein is an Associate Professor of Computer Science and STMicroelectronics Faculty Scholar at Stanford University.

Acknowledgements

We would like to thank our participants for their time and valuable feedback as well as the broader community of Discord moderators who have helped provide context for this line of work. We would also like to thank Mitchell Gordon, Michelle Lam, Joon Sung Park, Lindsay Popowski, Helena Vasconcelos, and members of the Internet Rules Lab at CU Boulder, as well as a number of other lab members and collaborators for additional feedback.

...

Disclosures: Michael Bernstein's lab, of which Joseph Seering is a member, received a $5,000 gift from Discord to help support a summer Research Assistant on another project. In addition to his role at Stanford, Joseph Seering was also a part-time consultant at Twitch starting after submitting the first draft of this paper and continuing through the publishing process, but he has no professional financial relationship with Discord.

Data Availability Statement

Not applicable

Funding Statement

This material is based on work supported by the National Science Foundation under Grant #2127309 to the Computing Research Association for the CIFellows Project. This work was also funded in part by NSF Grant #CCF-1918940, and by the Brown Institute for Media Innovation.

Ethical Standards

This research received approval from the Stanford University Institutional Review Board under protocol number 61409.

Keywords

Discord, LGBTQ+ communities, hate speech, harassment, moderation