New Media & Society
XX(X):1–28

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Moderator Engagement and Community Development in the Age of Algorithms

Joseph Seering[1], Tony Wang [1], Jina Yoon [2], and Geoff Kaufman [1]

Abstract

Online communities provide a forum for rich social interaction and identity development for billions of internet users worldwide. In order to manage these communities, platform owners have increasingly turned to commercial content moderation, which includes both the use of moderation algorithms and the employment of professional moderators, rather than user-driven moderation, to detect and respond to anti-normative behaviors such as harassment and spread of offensive content. We present findings from semi-structured interviews with 56 volunteer moderators of online communities across three platforms (Twitch, Reddit, and Facebook), from which we derived a generalized model categorizing the ways moderators engage with their communities and explaining how these communities develop as a result. This model contains three processes: being and becoming a moderator; moderation tasks, actions, and responses; and rules and community development. In this work, we describe how moderators contribute to the development of meaningful communities, both with and without algorithmic support.

Keywords

Online communities, Moderation; Governance; Social Networks; Platforms; Twitch; Reddit; Facebook

[1]Carnegie Mellon University, USA
[2]Brown University, USA

Corresponding author:
Joseph Seering, Human-Computer Interaction Institute, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA.
Email: jseering@andrew.cmu.edu

Introduction

Online social platforms are host to a wide variety of malicious behavior from spam and flaming to hate speech and extreme content. This is not a problem unique to online media, but it is one that has grown in importance as these platforms have become more ubiquitous. Recent attention to moderation in both research and public discourse has focused on company-driven removal of unwanted content at scale and the corresponding responsibilities of platforms (Klonick, 2017; Gillespie, 2018). For example, Roberts and Gillespie (Roberts, 2016; Gillespie, 2018) identify and discuss the use of many thousands of commercial content moderators whose job is to filter through an endless stream of content and remove what is deemed unacceptable on a given platform. While the politics and mechanics of content detection and removal strategies have taken center stage, the labor done by users to moderate their own communities has not received as much attention. Sites like Wikipedia, Reddit, Twitch (Lo, 2018), and Facebook Groups rely on their own users to do the vast majority of moderation work from the bottom-up, creating a significantly different dynamic than in spaces where moderation is driven top-down by company policy. User-driven moderation is an intensely social process that is core to community development.

The recent emphasis on scale follows a divergence in the research landscape on moderation as a whole. While early research focused on misbehavior as a group-level phenomenon in user-organized and user-managed online spaces such as Usenet newsgroups, Internet Relay Chat (IRC), and Multi-User Domains (MUDs) (Sternberg, 2012), more recent work has addressed misbehavior as a platform-level phenomenon following the rise of centrally-organized and managed online spaces like Facebook and Twitter (Crawford and Gillespie, 2016; Klonick, 2017; Gillespie, 2018). Despite the shifting focus, many self-governing online communities continue to thrive and, in some cases, are expanding rapidly. In this work, we detail a comprehensive model of how volunteer human moderators govern their communities in the age of algorithms.

We present the results of interviews with 56 moderators from three major social platforms: Facebook, Reddit, and Twitch. We identified complex social processes that drive who becomes a moderator in these spaces, how these users learn to moderate, how they deal with incidents, how they formulate rules and set norms, and how their communities evolve as a result. We found that moderators in these spaces feel a strong commitment to their communities, deriving personal meaning from guiding them and helping them develop. Rather than seeing misbehavior as something that could be "cleaned up" by algorithms or bans, many moderators choose to engage personally during incidents to set an example for future interactions.

We identify and describe three interconnected processes that drive moderation and governance in these communities. First, we describe the processes of becoming a moderator, including appointment into the role and development over time. Second, we detail the processes for handling incidents that arise both proactively and reactively. Finally, we describe the decision-making processes for modifying how communities are run. We focus on these three main themes because volunteer moderation is responsible for shaping the online experiences of billions of users distributed among many millions

of communities. Past research has explored pieces of these processes, usually on one specific platform. This systematic description of the full process of governance across diverse platforms can inform future research on the effectiveness of different strategies, guiding development of better tools that support, rather than supplant, the judgment of users.

Twitch, Reddit, and Facebook

The social dynamics of online communities have been explored from numerous angles. In this work, we focus on informal public spaces where people meet to share interests, converse, and build a community (Oldenburg, 1999), emphasizing the social nature of the spaces and the emergent community-building processes. Facebook Groups, Reddit subreddits, and Twitch channels (Hamilton et al., 2014) all meet these criteria, while network structures such as Twitter followers or Facebook friends currently do not. Though there are many online communities that match these descriptions, we chose Twitch, Reddit, and Facebook as three of the largest community-based sites with significantly different features and cultures. Facebook reports over 2 billion monthly active users[1], and Reddit[2] and Twitch[3] report hundreds of millions. While Twitch is much younger than Facebook and Reddit, it has been the subject of much research on community dynamics in the past five years (Hamilton et al., 2014; Wohn et al., 2018; Hilvert-Bruce et al., 2018).

Each of these platforms hosts different types of communities with different feature-structures. Reddit communities take the form of text-based discussion forums, where visibility of content is determined by voting (Massanari, 2017). Twitch communities are chatrooms built around interaction with a single specific user, the "streamer," who appears on a live video stream (Hamilton et al., 2014). All three platforms provide basic algorithmic tools for handling common misbehaviors. Reddit offers AutoModerator, a bot that proactively catches messages based on user-chosen settings for moderators to review later. Twitch's AutoMod functions similarly, though it requires more immediate attention due to the synchronicity of conversation on Twitch. Facebook's most commonly used algorithmic tool is its automatic flagging of group join requests from suspected spam accounts. However, despite the presence of these tools, the vast majority of moderation decisions are still made by users either manually or through independently developed bots.

In contrast to both Twitch and Reddit, which are platforms built to host communities, Facebook hosts Groups as a complement to the site's primary social network function. Facebook scales its content moderation by designing algorithms and employing thousands of commercial content moderators to tackle its massive network. (Roberts, 2016; Gillespie, 2018, p. 120-124).

Moderation and meaningful communities

Much work in the study of moderation has focused on the specific problems that occur and how they are handled, such as the vexing question of how to deal with "trolls" that

plague an online community (Herring et al., 2002). Early work analyzed misbehaviors that appeared in communities that were dominantly user-run, such as Usenet newsgroups, MUDs and, more recently, Wikipedia. In a review of online misbehavior in the early social web, Sternberg identifies an "Infamous Triad" of flaming, spamming, and virtual rape (Sternberg, 2012, p. 77-85). Spam and broad incivility remain problematic on all types of platforms, and more recent work on online harassment has explored targeted attacks often focused against particular identity groups (Fox and Tang, 2017).

While much work has focused on how specific behaviors are handled, it remains an open question how moderators differentiate between the wide variety of behaviors that happen across different platforms, and how these strategies evolve over time. In Herring et al.'s aforementioned work, members of the community engaged with the "troll" through debate or insults, called for his removal or for other members to ignore him, and started conversations to try to come to consensus about rules and norms for the space. Eventually, a moderator took independent action to remove the offender. Herring et al.'s work does not elaborate on the moderators' thought processes or how they might apply to other types of behaviors, and literature on how moderators attempt to reform troublesome users is rare. Very little is known about what happens to individual users following disciplinary action, particularly if the action was intended to help them reform, a gap which the present work aimed to fill.

Though user and group-level strategies for moderation continue to be studied (Kiesler et al., 2012; Seering et al., 2017; Jhaver et al., 2018), much recent work has begun to study misbehavior as a network-level phenomenon that can be dealt with using a top-down approach. This represents an implicit shift from managing misbehavior to filtering content. The increasing application of machine learning methods to social computing problems offers a solution for detecting a wide variety of negative behaviors at scale. (Gillespie, 2018, p. 52-63) Albeit imperfect and often difficult to define, these approaches are important as platforms continue to seek scalable moderation methods. To supplement these algorithms, major platforms also hire thousands of commercial content moderators to sort through suspect content (Roberts, 2016; Gillespie, 2018). These top-down approaches emphasizing commercial content moderation lead to several additional questions: How do user-moderators make use of or interact with these tools? When do users want or not want to use the algorithmic tools made available to them? How does reliance on these tools affect how communities evolve? Following (Fiesler et al., 2018), we know that volunteer moderators develop complex, community-specific rules as their communities evolve, but the literature lacks a generalizable, cross-platform model for how these decisions are made over the life cycle of a community.

The overarching theme emerging from existing community moderation literature is that communities evolve over time as a result of rule-breaking, rule-making, and rule-enforcement (Sternberg, 2012, p. 158-169). Rule-breakers often include malicious outsiders, spammers, or trolls but sometimes also regular users who misunderstand rules or get carried away. It is important to note that, while these approaches are content-focused, users are not passive participants in the moderation cycle; they actively monitor and react to both algorithmic and human moderation decisions to gauge what is

appropriate or not. Some even do this to evade detection for content that they know will violate the rules (Friedman and Resnick, 2001; Gerrard, 2018).

Grimmelman defines moderation as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse." (Grimmelmann, 2015, p. 52). This definition matches well with our findings in this study; our interviewees moderated communities based around computer games, board games, art, memes, cars, sports, pets, and learning. They acted as arbiters, governors, community managers, teachers, role models, curators, and enforcers. Modern online communities look much like their earlier incarnations both in the misbehaviors they handle and the strategies they use to guide the community, but they exist in a new context: working alongside - and sometimes at odds with - platform-driven algorithmic moderation. The present work aims to provide an in-depth examination of user-driven moderation to shed new light on the development and decision making processes exhibited by volunteer moderators across diverse platforms.

Methods

We performed 56 semi-structured interviews from Fall 2016 through Spring 2018. We began with 20 semi-structured interviews of Twitch moderators, using several snowballs for recruitment. See Appendix A for the interview protocol. We directly messaged active streamers and moderators with different backgrounds and interviewed up to two of their connections. In order to identify a broad set of experiences, we recruited moderators from communities built by streamers of different genders, nationalities, and sexual orientations. Based on our results, we added a section on relationship with platform administrators, employees of the respective companies, to our interview protocol, which we detail in the related section. See Appendix B for a list of interviewees and community characteristics. All interviewees for this study were paid $15 for participation. Interviews lasted between 25 and 55 minutes, with variance according to number of communities moderated and depth of engagement within communities. All interviews took place remotely via Skype, Discord, or Messenger voice calls, with audio recorded for later transcription by the researchers. In two cases, we were also sent documents by interviewees related to their roles as moderators.

In the second phase of this project we interviewed 21 Reddit moderators[4]. In recruitment for this sample, we messaged moderators from small (3,000-10,000 subscribers) and large (200,000+ subscribers) subreddits. These included, for example, subreddits focused on shared interests like cars or games or pets. These informal communities based around shared interests are important to understand because of their impact on user identity and development.

Third, we interviewed 15 Facebook Group moderators[5] during Fall 2017 through Spring 2018. We focused on the same types of groups as on Reddit, using keywords related to the subreddits from which we had previously interviewed moderators (e.g., "cars" and "memes"), with the goal of finding comparable communities on all three platforms. We selected Facebook groups of various sizes (500-70,000 members) and messaged recently active moderators for interviews.

Once all interviews were completed and transcribed, we "winnowed" the text into chunks for coding following the procedure established in Creswell (Creswell, 2013, p. 86-89, 184-185). Each chunk contained a single idea and varied in length from a few words to three sentences. Our final dataset contained 1,877 chunks of text. One rater assigned codes to a subset of these chunks. First, low-level themes were identified, and then these themes were abstracted to higher levels of generality to produce a comprehensive codebook.

To ensure inter-rater reliability, we calculated Cohen's Kappa statistics using two coders working independently. We began by calculating reliability for assignment of chunks to each of the three top-level processes, and then proceeded to calculate reliability of assignment of codes for steps within each of these processes. Initial inter-rater reliability statistics were low ($kappa < 0.60$), so the codebook was revised to account for disagreements. Additional subsets of chunks were coded to re-compute with the updated codebook. Cohen's Kappa statistics for the final codebook presented here ranged from 0.70 to 0.89, indicating moderate to very strong agreement. Following these tests, two researchers independently coded each of the 1,877 chunks. Disagreements were resolved through discussion.

This sample is not a random sample of moderators on these sites. Female and LGBTQ+ moderators are intentionally over-represented in our sample, especially on Twitch, in order to gather a more diverse set of experiences. Experiences of underrepresented moderators are especially important to capture on Twitch because the video-based structure of the platform makes their identity more salient to community members, and because of the prevalence of sexism in game-related media (Fox and Tang, 2017).

The Moderator Engagement Model of Community Development

We present a model of three primary processes in which moderators are engaged, each contributing to community development over a longer period of time. These processes are not necessarily sequential but rather comprise the many often simultaneous duties of a moderator. (1) Over the course of weeks or months, new moderators are chosen, learn through daily interactions, and develop a moderation philosophy. (2) Moderators interact on a daily basis with users and make individual short term decisions about specific incidents, ranging from warnings to light penalties and eventually to bans if necessary. (3) Finally, throughout the life cycle of the community, moderators make important decisions about policies that impact how the community evolves, usually in reaction to problems that emerge. These decisions are often made without substantial feedback from non-moderators.

These three processes interact fluidly and lead to community evolution over time. The following examples illustrate paths through and between the three processes that we observed in our interviews:

- F7 started moderating a professional Facebook Group after volunteering to help. The first few times F7 took action as a moderator, F7 would ask more experienced moderators if the action they were about to take was reasonable. After building

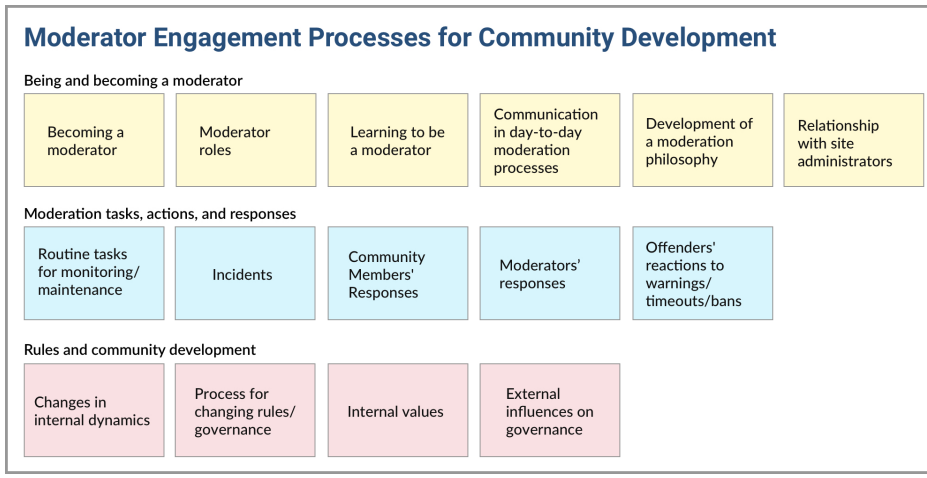

Figure 1. Moderator Engagement Model of Community Development

trust with the moderation team, F7 started moderating independently. Recently, F7's group had troublemakers that responded with hostility to warnings. F7 was part of a team of moderators that made the final decision to ban these members and took responsibility for deleting inappropriate comments. Here, becoming and learning how to be a moderator are intertwined with community responses to bad behavior, strategies for handling incidents, and violators' responses.

- R21 has been moderating a sports subreddit for three years. Instead of being given formal guidelines, R21 was told by more senior moderators to moderate however R21 feels is best for the subreddit. The moderation team is particularly wary of violating a Reddit rule against posting personal information, so R21 spends time warning users about this. R21 has worked with Reddit administrators before, so when a user repeatedly violated the subreddit's rules by making new accounts after being banned, R21 reached out to them for help. In this example, specific incidents led to community-specific rule enforcement styles and eventual engagement with platform administrators.
- T18 streams on Twitch to an average of several hundred concurrent viewers. T18 has no hard-line rules about moderation. Instead, T18 focuses on engaging viewers who share different opinions and only removes spam bots. Though Twitch has added new moderation tools in recent years, T18 feels that these tools have no place in his community because moderation tools stifle conversation and inhibit safe and productive community growth. In this third example, a particular moderation philosophy adheres to social rather than technical forms of moderation, which are intended to help the community grow in a particular way.

Being and becoming a moderator

The ways in which individuals become moderators are similar across platforms and are largely driven by discussions within moderating teams. Platform administrators are occasionally involved, but these instances are rare and often opaque to the moderators. The steps shown in column one of Figure 2 match steps from the overall process diagram, Figure 1. The second column shows themes and variants within each step. See Appendix C for counts of each code.

Becoming a moderator There are five common paths to becoming a moderator. Three rely on social presence and existing connections and two require demonstrable qualifications. Moderators are most commonly selected for the position because they were standout members of the community; head moderators tend to look for members who understand the community's values, have the maturity to set an example, and can enforce the rules appropriately.

> "Mostly my [moderators] come from people who have been members of my community for a long time, who have intelligent opinions about things, who have shown me that they can be reasonable about things that are even a little bit difficult sometimes." - T19

Users could also become moderators because they had an existing relationship with current moderators of a community. This practice is more common on Facebook than either of the other two sites, likely because Facebook is a social network based on real-life relationships and allows for "friending." Female moderators in smaller Twitch communities noted the value of having friends moderate for them, particularly ones who understood the sexism they faced. This matches recent work on "friendsourced moderation" (Mahar et al., 2018).

Finally, many moderators noted that it was important to have a team that could cover all hours of the day, leading them to select moderators in different time zones. This was less common on Twitch channels, possibly because of the time-bounded nature of conversation on Twitch.

Role differentiation We found relatively little evidence of specific role-differentiation in our interviews. Only a handful of moderators were selected solely for their design or technical skills. Despite relatively low frequency of appointment of moderators for specific types of tasks, e.g., graphic design or technical support, moderators frequently discerned levels of authority.

> "We have a very structured hierarchy where our head mod who created the subreddit is like the president." - R2

Though differentiated authority was reported across all three platforms, the way it manifested depended on the platform's design. On Twitch, role differentiation is explicit in every community: the channel owner (usually the streamer) has ultimate authority in their own community. Facebook offers different levels of power through the "administrator" and "moderator" designations. Reddit has less obvious titles, but

1. BEING AND BECOMING A MODERATOR	
Step	**Theme**
Becoming a moderator	Friend, family member, or connection
	Recognized from other moderating experience
	Stand-out member of the community
	Availability at important times of day
	Volunteered or applied to become a moderator
Role differentiation	No different roles
	There is a head mod and/or hierarchy
Learning to be a moderator	Discussion or instructions
	Implicit understanding from being in community
	Learning by doing
Communication between moderators	Discussion about moderation decisions
	External platforms are used for communication
	Internal platforms are used for communication
	Off-topic or social conversations
	There is little or no communication
Development of a moderation philosophy	Valuing direct engagement
	Hands-off approach
	Maintaining a neutral stance
	Moderators as group “police”
Relationship with site administrators	Little or no engagement
	Work together to address problems

Figure 2. Steps and variants in Being and Becoming a Moderator process [κ = 0.89]

more senior moderators can add new moderators with limited permissions. Each of these design decisions facilitates a slightly different type of status differentiation.

Learning to be a moderator Formal education of new moderators was strikingly rare on both Twitch and Reddit. More than two thirds of moderators learned based on a combination of understanding the community's values and simply learning by doing.

> "They [other moderators] just said do whatever you feel makes the subreddit better so I've been rolling with that since I became a mod." - R21

Fewer than a third of our interviewees had any formal onboarding conversation about what was expected of them, and even fewer were given formal guidelines. Of the three platforms, these conversations were most common on Facebook; moderation processes are typically invisible to casual Facebook users who are not involved in Groups (Myers-West, 2018), so Facebook users may have simply needed more of an introduction to them:

> "If the account is less than a year or so old, be wary. Many of these are socks [sockpuppets], you can especially tell if no real photographs are used. They could be bots too, you can tell by the content shared" - From onboarding document shared by F13

Communication between moderators and Development of a moderation philosophy Though different platform moderators communicated on different tools, the topics discussed were fairly similar. The most common conversations were about specific incidents, seeking advice or opinions on the best response, or notifying other moderators about an action taken.

> "We discussed it and we were all on the same page that a couple of people had just crossed the line too for and we went ahead and removed them." - F7

This was less common on Twitch. Due to the synchronous nature of conversations on Twitch, moderation decisions need to happen immediately. However, Twitch moderators communicated in other ways, including "mod meetings" where incidents and rules were discussed outside of streaming hours. Some moderators indicated that there was rarely any communication at all, and though they did not usually see this as a problem, they were typically from groups that were either homogeneous or stagnant. A lack of conversation often indicated a lack of growth. One Twitch moderator (T18) noted that this was the reason he avoided algorithmically-driven moderation; ongoing discussions about acceptable behaviors helped both him and his community grow, and automating these decisions could remove opportunities to have these conversations.

Over time, moderators developed a philosophy through conversations with other moderators and engagement with users. Four different philosophies emerged in these interviews. A number of moderators stated that they felt it was important for them to be very present in the community by setting an example and engaging directly. Other moderators saw themselves as "police", tasked with identifying and punishing offenders

in order to keep the community civil. A third group of moderators sought to maintain a neutral perspective. They felt suspected bias could affect their credibility or hinder open discussion. They stepped in when conversations got out of hand, but did not take stances themselves. And finally, a small group of moderators believed in a hands-off approach, interfering as little as possible even when discussions devolved into personal insults:

> "If it seems like people are disagreeing, even if it's getting really really heated and personal, [...] if they're responding to actual content in the conversation rather than just throwing slurs at each other and posting memes and empty agitation and responses, then that's good to me." - F11

This fourth philosophy was associated with a strongly anti-"censorship" ethos (Massanari, 2017), with the idea that debate and discussion naturally surfaces better ideas and greater understanding.

Relationship with site administrators In the first round of interviews with Twitch moderators, we occasionally heard about interactions with Twitch employees:

> "I feel like, at Twitch's level, it's more about managing streamers, like I manage my community and I expect every other Twitch streamer to do the same [but] I think that Twitch's job is to manage streamers, to control that space, so if a streamer is being disruptive, and harmful to the community at large, it's Twitch's job to manage them." - T18

This suggests a clear division in labor between community moderators and platforms; volunteer user moderators feel that it is their job (and often their right) to manage their communities and that platform employees should only intervene if things go very wrong. In response to this emerging sentiment, we added questions to our second round of interviews to capture moderators' relationships with admins.

Prior work (Matias, 2016) has explored moderators' relationships with platform administrators, noting moderators' need to maintain face. We found few examples of this; most of our interviewees felt no connection to administrators and believed their communities would never host behaviors that would attract administrator attention.

> "I don't know if they ever review us. I'd be surprised if they knew we existed to be honest." - R17

Virtually all Facebook and Reddit moderator interviewees reported little to no engagement with platform admins and also a general uncertainty about how admins decided where to direct their attention. Some Reddit moderators noted that they did occasionally engage with platform admins, but only to get help on technical problems they could not resolve, such as users who repeatedly create new accounts to circumvent bans. The differences between these findings and prior findings likely have to do with the focus of the aforementioned work. Matias focused predominantly on large and fairly active subreddits, while our interviews also captured the dynamics of smaller communities that were unlikely to cause administrators trouble.

Beyond these relationships, several moderators reported "mod burnout" in which moderators became exhausted by the amount of work and exposure to offensive content. Several moderators recalled traumatic experiences like threats and harassment, some cases of which even followed them offline. To many moderators, moderation is equivalent to a second job where they work for the benefit of the platform and are rewarded only with the satisfaction of helping shape a community that they care about (Matias, 2016).

Broadly, the process of becoming a moderator moves from initial appointment based on behavior or connections to their learning and development process. Communication between moderators about their experiences is core to this development. This process occasionally includes interaction with platform admins, but our work reveals that moderators' attention was primarily internal. They felt both the capability and the right to manage their own communities without interference.

Moderation tasks, actions, and responses

Though a moderator's development primarily happened through communication, much of their time was spent dealing with misbehavior, as shown in Figure 3. Monitoring a space by identifying and responding to offenses is a complex social process; a single incident often involves several moderators and community members. While moderators identified some useful proactive tools such as filters that held posts for review, most cases require them to react based on the community's standards and the offender's perceived intentions. Offenses that result from misunderstandings or brief losses of composure are often dismissed with warnings or temporary restrictions, whereas intentional or egregious violations are more likely to warrant severe penalties like expulsion from the community.

Routine tasks for monitoring/maintenance The asynchronous nature of Facebook Groups and Reddit encourages moderators to develop proactive strategies for preventing misbehavior. This reduces the amount of work they have when responding to incidents after they happen. Almost all Facebook Groups interviewees reported that they spend a significant amount of time reviewing join requests, a feature that does not exist on either Twitch or public subreddits. Join requests give moderators the ability to accept or deny prospective members, which moderators reported was effective at reducing spam but created an additional workload for them. While most groups only filtered to prevent spam accounts, some groups used filtering to curate members that had similar values, interests, or political orientations. Facebook Groups moderators also spent significant amounts of time monitoring potentially controversial threads and stepping in to preclude conflicts. Prior work has shown that there are detectable patterns of behavior that precede misbehavior (Liu et al., 2018), and Facebook moderators are likely attuned to these.

A number of moderators, particularly on Twitch and Facebook, felt it was their duty to engage regularly with their community. Some Facebook moderators contributed content on a regular basis to keep discussion flowing, and Twitch moderators regularly welcomed new members and actively answered questions. Some framed these behaviors as setting a positive example. Prior work has found that users imitate authority figures' behavior online at significant rates, and moderators took advantage of this (Seering et al., 2017). Reddit moderators were generally more hands-off in this sense.

2. MODERATION TASKS, ACTIONS, AND RESPONSES

Step	Theme
Routine tasks for monitoring/maintenance	Approving new members
	Contributing to the discussion
	Keeping the space "clean" or managing potential conflicts
Incidents	Disruptive behaviors
	General incivility
	Targeted attacks
Community Members' Responses	Critiquing offenders, explaining rules, defending community
	Flagging or reporting content
Moderators' Responses	Banning, timing out, or muting users, removing content
	Explaining to users why they were punished
	Use of tools beyond bans/timeouts for moderation
	Warning offenders
Offenders' reactions to warnings/timeouts/bans	Escalate behavior or resist
	No reaction
	Reform or apologize
	Seek clarification or request review

Figure 3. Steps and variants in Moderation Tasks, Actions, and Responses process [κ = 0.70]

Incidents We identified three categories of misbehavior commonly reported by moderators. First, "disruptive behaviors" included advertisements, spam, repeated nonsense, and malicious links. Though these types of incidents were irritating, moderators did not feel they threatened the community. Another category was "targeted attacks", which were often directed at underrepresented users such as women on Twitch or Reddit. In most spaces, these identity-based attacks were treated severely but at different levels depending on the specific type of group targeted. Racism was explicitly prohibited in communities more often than sexism, homophobia, or ableism.

We also identified a third category, which we term "general incivility". This includes general rudeness, impoliteness, and social faux pas. As communities grew, moderators encountered a greater frequency of incivility, often from users unfamiliar with the rules. Communities in their earlier stages are likely populated by individuals who share common values, but as they grow, they encounter users who do not share this understanding. Many moderators discussed how rapid growth leads to moderation challenges, and often assumed that users who misbehaved were outsiders or newcomers.

Community members' responses Though moderators have the final say in how their communities are run, they were quick to point out that general community members do an enormous amount of valuable moderation work by themselves, both socially and through site features. On Reddit and Facebook, the most common community response to misbehavior was to flag it for moderators to review. Moderators on these platforms said that, though sometimes abused, these reports were the easiest way to find misbehavior within the large volume of content produced. The synchronicity of Twitch conversations makes flagging more difficult, and the feature only sends the content to site administrators rather than local moderators. Twitch users often instead post messages in the channel immediately to ask moderators to deal with a disruptive user.

General community members on all three platforms often verbally critiqued or educated rule-breakers on proper behavior-

> "It's funny, cause a lot of the time I don't even have to say anything because my community does the shunning themselves." - T11

Moderators reported that verbal rebukes from the general community were usually helpful, and some Twitch moderators even found them emotionally validating.

Moderators' responses When proactive approaches and general community rebukes failed, moderators stepped in to address misbehavior directly. In general, moderators' responses to offenses began as light, verbal warnings and escalated into increasingly technical restrictions after repeated occurrences such as stronger warnings with a temporary mute and, eventually, a permanent ban. Exceptions included spam, egregious content, or suspected non-human accounts - these offenses typically warranted immediate bans. Virtually all moderators on Twitch and Facebook reported warnings as the first step in many cases, especially if offenses were mild and unintentional. As in Slovak (Slovak et al., 2018), moderators saw themselves as arbiters of the rules but also as teachers helping users learn how to behave.

> "We reply to the post with a warning. I think it's good to publicly give warning to show the community that we're taking action and also as a warning to other people so that they also know that this behavior isn't accepted." - R1

Nearly all moderators mentioned using timeouts, bans, or equivalents, though eagerness to use them varied. Communities with more laissez-faire ideologies used these only for egregious offenses, while communities intended to be safe spaces were usually

quicker to use them. While Twitch moderators relied extensively on warnings, they rarely issued explanations of punishments after the fact unless privately contacted by a user. This is likely because of the synchronicity of conversation on Twitch. On the other hand, Facebook and Reddit moderators frequently explained why they punished users or removed content, especially for posts that were left "hidden" until changes to the content were made to fit moderator approval.

There were three main platform-provided, algorithmically-based moderation tools used by moderators in this study. Moderators on Reddit relied substantially on the built-in AutoModerator to parse, flag, or remove suspicious posts. Despite its utility, AutoModerator sometimes created additional work for moderators because they had to manually approve posts mistakenly "caught" by the bot. Twitch moderators relied less on site-provided automated tools with the exception of the emerging "AutoMod" tool, [6] and Facebook Groups moderators relied the least on automated tools to parse posts for them, though some made use of built-in filters in the user approval process where potential spam accounts were marked for review.

Many Twitch and Reddit moderators looked beyond site-provided tools and used free user-developed bots; some even created their own. Most of these indepdently managed tools were simple filters that flagged custom words; moderators preferred to engage verbally with human offenders in more nuanced situations. Beyond bots, the most commonly used tools were chat logs, post histories, and ban logs. Facebook Groups moderators were least likely to integrate custom tools because of third-party developer restrictions on the site.

Moderators we interviewed were happy to have tools that deal with the most obviously unwanted content, such as links to malware or pornography, but they have a strong preference to make the hard decisions themselves. This stands in contrast to prior work on preferences for algorithms in other contexts. When exploring users' preferences for algorithmically-driven tools in medical decision-making processes, Yang (Yang, 2017) found that doctors felt no need for these tools for most of the easy decisions they made because they felt confident in their decision-making ability, preferring support instead on the harder decisions. The difference between these cases likely results from the importance of continuously-evolving community values in decisions made by moderators. They noted both the importance of their ability to make context-specific judgments and also the impact on the community's development that their decisions have as justification for reserving these decisions for human judgment.

Offenders' reactions to warnings/timeouts/bans Offenders react to punishments in one of four ways. The most common response was actually *no* response. These offenders might not have cared enough or might have expected to be punished. Spambots were also unlikely to respond. In fewer cases, however, some offenders continued or worsened their behavior. Moderators escalated punishments accordingly,which might ultimately lead to a ban. Sometimes, even after getting banned, persistent offenders continued to harass moderators in other ways. These reactions could reach dangerous levels, particularly when the moderator involved was personally identifiable:

> "One time a person who I had banned went on to the Facebook page of the place where I work and said some incredibly rude and obscene things about me." - F6

Alternatively, many moderators reported that some users actually responded well to warnings or light penalties by reforming or apologizing soon after. Moderators felt that these users probably did not understand the rules or had just gotten carried away.

> "My favorite is when people actually come to me and say, I said this, I didn't realize that that was going to be upsetting to people and I apologize, I won't do it again. Can I be unbanned?' And I love unbanning people for that." - T20

Offenders also commonly requested clarification about their punishment. Some questioned the rules or pointed to other examples (e.g., "why did I get banned and he didn't?"). Making nuanced decisions about punishment on a case-by-case basis was one of the greatest challenges moderators discussed.

Top-down approaches in commercial content moderation are designed to minimize subjectivity in moderation work (Gillespie, 2018, p. 111-114). According to our interviewees, however, it is this very subjectivity that helps communities develop over time. The values that moderators brought to their commmunities and the ways that these values changed as a result of interactions with users were core to community growth. In cases where decisions were made by algorithms rather than moderators, these moderators might not have the same opportunities to grow.

Rules and community development

Communities and their rules develop over time through reactions to short-term events or transitions, as shown in Figure 4. As rules and norms develop, they drive subsequent moderation decisions that shape community identity. These decisions are almost exclusively made by moderators, either by an executive "head" moderator or a group consensus. Occasionally, moderators solicited feedback from their communities, but this was rare. General community members are rarely given a say in the final outcome.

Changes in internal dynamics Virtually all rule changes were made in response to unexpected incidents either gradually over time or suddenly following a specific incident.

> "If the rule is there, it's because somebody broke it." - F15

Moderators considered changes either when users began to misbehave in a way that was not expected or when implied norms needed to be made more explicit. The most common precursor to such incidents was a sudden diversification of the community. This might happen when outsiders who hold a different set of values join the community, or when malicious users target the community with the intention to disrupt it. For example, Reddit automatically gives visibility to posts that receive particularly high user vote scores, which can introduce new users to a community. Twitch both selectively highlights

3. RULES AND COMMUNITY DEVELOPMENT	
Step	**Theme**
Changes in internal dynamics	Community evolves and/or grows over time
	Issues or problems arise
	Temporary special situations
Process for changing rules	Community input
	Discussion among mods
	Executive decision
External influences	Site rules
Internal influences	Personal values

Figure 4. Steps and variants in Rules and Community Development process [κ = 0.85]

communities on its front page and allows communities to direct their members toward other communities via "raids".

Rules might also change in response to unusual internal or external events. One Facebook group moderator noted that large volumes of posts in response to political news (e.g. North Korean nuclear tests) spurred moderators to temporarily restrict this type of content in order to prevent their communities from being dominated by a single topic. Similarly on Reddit, some moderators noted how rules changed in response to the growing popularity of "meme" submissions that detracted attention from more discussion-based content. A Twitch moderator reported that rules changed when they had a special guest on stream who might be the target of particular types of attacks such as gender-based harassment; moderators were intentionally stricter during these events.

Process for changing rules As communities matured, moderators gained a clearer vision of what they wanted within their communities. Moderators reported that their initial community was made up of people they understood or identified with, but as the community grew, so did the frequency of misunderstandings. Slower community growth was much easier for moderators to manage than sudden influxes of new users. Rapid growth or inconsistent enforcement led to more chaotic communities.

> "As subreddits get bigger there's stuff you didn't even think about and you have to make rules for, like hate speech, racism, and t-shirt company spam."
> - R20

We noted three major common processes for rule change, all of which varied in who had input into decisions. In communities with a clear hierarchy, head moderators often made final decisions and sometimes even announced changes without asking for feedback. Despite the lack of involvement, most moderators in these communities accepted this as a legitimate process. In communities with less structured hierarchies, the most common process for changing rules was an open discussion among moderators about the change and how it should be communicated to the community. This process varied in formality from informal requests for thoughts to a specified period for debate (e.g., two weeks).

> "There would be a proposal submitted over modmail to the mod team as to what the change might look like, and then the team would provide their input." - R7

A small number of moderators, mostly from Reddit, described processes for getting community input on changes or issues. One method was a survey deployed to the general community, often generated using Google Forms. Another was to allow comments on a post with proposed rules changes to elicit feedback. There were also "meta" subreddits for discussing community logistics rather than content. However, several Facebook and Reddit moderators stated that they actually found it easier to *avoid* transparency in rule changes and enforcement because of the conflicts that arise from announcing decisions that general community members often would not notice otherwise. While community input was occasionally considered, major decisions were made exclusively by moderators and community members could not vote. In only one case on Reddit did changes result from collective pressure from the general community:

> "If we see a lot of people complaining about a rule we re-evaluate it. Back before when [the community] was a lot smaller it used to be more lenient about generic posts, like heres a photo of me at the game or heres a photo of me with a player but we kind of put a stop to those just because [the community felt] it wasn't as interesting to discussion."

External and internal influences We identified two other sources of influence in making these decisions: external influences (such as sitewide policies) and internal influences (such as personal values). Influence of site rules and content policies on community rules was rarely present, likely due to the vague and distant nature of these policies (Gillespie, 2010; Pater et al., 2016). Reddit's official policies did serve as a minimum standard for behavior in subreddits, as failure to comply could lead to shutdown of the community. While Reddit moderators had little to no contact with site admins, they did note that specific policies impacted how they moderated, with the most common being Reddit's policy against revealing personal information. Moderators on Twitch and Facebook usually did not pay attention to content policies and, in many cases, did not know what they were.

A smaller number of moderators mentioned that their experiences in other communities - whether as moderators or general community members - influenced

their philosophies about moderation in their current groups. For example, Facebook moderators mentioned that communities were sometime split-off from other communities with a subset of their membership. Though Fiesler et al. (Fiesler et al., 2018) found that only 3% of rules on Reddit appear verbatim in multiple subreddits, it is plausible that moderators for 'split-off' subreddits took their prior experiences as a baseline but re-wrote the rules for the new context.

Conclusion

This article outlines three processes through which moderator engagement guides the development of online communities: becoming and developing as a moderator, handling misbehavior, and developing rules for the community. It contributes to the growing discussion surrounding management of behavior in online spaces by documenting the social nuances of moderation that disappear when moderation is delegated to commercial content moderators or automated algorithms. These findings emphasize the need for a closer look at social, user-driven models of moderation. The communities described by just the 56 moderators in this study are comprised of more than five million total members, many of whom find valuable connections, relationships, and meaning in these spaces. While recent work has suggested that social media companies may be the "New Governors" of the digital age (Klonick, 2017), it is important to remember that this centralized aggregation of power is not necessary for meaningful online socialization to occur. Users can be very effective at self-governing when given the tools to do so, and this experience in itself can be meaningful.

Both community-based moderation and commercial content moderation have clear drawbacks. As Gillespie (Gillespie, 2018) notes, there are a variety of challenges in commercial content moderation that platforms must address if they are to be the "custodians" of the public sphere. They must balance intervention with protection of users' rights to speak; they must make moderation decisions constantly; and they must maintain the appearance of fairness and objectivity. Though volunteer user moderators also sometimes struggle with transparency and fairness, they are much better equipped to understand the context surrounding issues in their communities. Moderators engage personally in dealing with a variety of nuanced problems, guide conversation in positive directions, and are a regular, stable presence in their communities. Moderation algorithms and commercial content moderators are unlikely to be able to contribute to these community attributes in the same way.

> "I see myself more as a gardener kind of mod so to speak. So I'm very active, planting new posts and also removing the weeds so any posts or comments that are very negative and very damaging to the community, I would want to remove." - R1

Prior research has explored ways to make commercial content moderation more effective because of its presumed scalability, but it is important to remember that online communities have been self-governing imperfectly but effectively since the beginning of the social web. Many of the challenges in scaling moderation, such as inability to consider

context of each potential violation, come from decisions to structure social platforms like Facebook and Twitter as networks rather than a series of self-governing communities. Even within Facebook itself, we find that characteristics of user-driven moderation within Groups are largely consistent with the characteristics within Twitch and Reddit communities; Facebook Groups offer an opportunity to test what a more community-based Facebook might look like and whether this model might be able to avoid some of the pitfalls of Facebook's default commercial content moderation approach.

There are many ways to balance algorithmic and user-driven models of governance, and each implementation has different implications for communities online. The concept of a "community" on Twitter is very different from on Reddit, in part because of where control over what is acceptable is situated. User-governed spaces, for example, may be more sensitive to local context and better able to support users personally in learning, discussing, and developing, but allowing users full control can create safe spaces for extremist communities and hate groups to develop and enforce their own norms. Future research should treat moderation as a balance between platform and user-driven governance, incorporating a focus on user agency that has been less prominent in recent work. There are also questions to be asked about the ethics of building a platform on top of user labor; several of our interviewees mentioned that their job was rewarding but exhausting, and two mentioned that they wished that the platform recognized their efforts. However, nearly all of our interviewees found their work as a moderator personally rewarding, and none expressed indication that they felt stuck in a position that they would prefer to leave.

We offer four additional considerations for future design in light of these findings. First, platforms should work to develop and improve tools that allow moderators to focus their attention where it is needed. For example, Facebook and Reddit moderators currently rely extensively on flags, but predictive suggestions for threads or discussions that might soon devolve could be useful as a complement to flags (Liu et al., 2018). Second, platforms should consider features that encourage positive behaviors in meaningful ways. While tools for dealing with misbehavior are common, tools for encouraging meaningfulness are limited at best (e.g., Reddit gold, Facebook reactions). Third, platforms might consider developing features that allow all users to get involved in self-governance if they so choose. Moderators in our study made decisions by executive fiat usually without community input, and could not be removed from their positions except in some cases by other moderators. This model of governance has been common across online communities since the early social web, but other models of community governance may be possible. Finally, platforms should consider where scaffolded user-driven moderation might serve communities better than algorithmic or company-driven moderation, and how social features might be designed to facilitate user-driven moderation. For example, when is a network a better design choice than a set of communities, and vice versa? What would Twitter and Facebook look like if they were structured primarily around communities rather than networks?

Meaningfulness in online spaces emerges from nuanced social interactions, both positive and negative, and volunteer moderators are at the core of these interactions. Through this analysis, we document how moderators help such communities grow,

evolve, and become more meaningful for their members as they work through the challenges that come from engaging with new media. We find that, while moderators did in certain cases make use of algorithmic moderation tools, they always sought to give the important parts of their jobs careful human attention and contextually-informed judgment. Future tools should support moderators in finding time and energy to focus on the tasks that they find meaningful and that help their communities grow.

Acknowledgements

We would like to thank Casey Fiesler, Jessica Hammer, and Laura Dabbish for extensive feedback on drafts, as well as Kat Lo for feedback on the research design and Greg MacDonough and April Sperry for editing support. We would also like to thank Neel Tiwary for support with interviews. Finally, we would like to thank our interviewees for taking the time to share their experiences with us.

Notes

1. *Facebook's grim forecast: privacy push will erode profits for years.* Reuters. Retrieved from https://www.reuters.com/article/us-facebook-results/facebook-misses-estimates-on-monthly-active-users-idUSKBN1KF2U5
2. *The conversation starts here*. Retrieved from https://www.redditinc.com/
3. *Audience*. Retrieved from http://twitchadvertising.tv/audience/
4. One Reddit moderator interview, R14, could not be transcribed due to audio issues, so is not included in counts in Appendix C
5. Our sample included both what Facebook calls "moderators" and "admins", both of which are types of moderators as defined by (Grimmelmann, 2015, p. 42) with the latter having more permissions. For the sake of simplicity we refer to both as "moderators" here and use the term "platform administrators" to refer exclusively to employees of the respective companies.
6. Note that our interviews took place largely prior to the widespread proliferation of Twitch's "AutoMod"

References

Crawford K and Gillespie T (2016) What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society* 18(3): 410–428.

Creswell JW (2013) *Qualitative Inquiry and Research Design: Choosing Among Five Traditions*. Thousand Oaks, CA: Sage.

Fiesler C, Jiang J, McCann J, Frye K and Brubaker J (2018) Reddit rules! characterizing an ecosystem of governance. In: *International AAAI Conference on Web and Social Media*.

Fox J and Tang WY (2017) Women's experiences with general and sexual harassment in online video games: Rumination, organizational responsiveness, withdrawal, and coping strategies. *New Media & Society* 19(8): 1290–1307.

Friedman EJ and Resnick P (2001) The Social Cost of Cheap Pseudonyms. *Journal of Economics & Management Strategy* 10(2): 173–199.

Gerrard Y (2018) Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society* 20(12): 4492–4511.

Gillespie T (2010) The politics of 'platforms'. *New Media & Society* 12(3): 347–364.

Gillespie T (2018) *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*. Yale University Press.

Grimmelmann J (2015) The Virtues of Moderation. *Yale Journal of Law and Technology* 17(1).

Hamilton WA, Garretson O and Kerne A (2014) Streaming on twitch: Fostering participatory communities of play within live mixed media. In: *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14. New York, NY, USA: ACM, pp. 1315–1324.

Herring S, Job-Sluder K, Scheckler R and Barab S (2002) Searching for Safety Online: Managing "Trolling" in a Feminist Forum. *The Information Society* 18(5): 371–384.

Hilvert-Bruce Z, Neill JT, Sjöblom M and Hamari J (2018) Social motivations of live-streaming viewer engagement on Twitch. *Computers in Human Behavior* 84: 58–67.

Jhaver S, Ghoshal S, Bruckman A and Gilbert E (2018) Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25(2): 12.

Kiesler S, Kraut R, Resnick P and Kittur A (2012) Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design. MIT Press, Cambridge, MA* .

Klonick K (2017) The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review* 131: 1598.

Liu P, Guberman J, Hemphill L and Culotta A (2018) Forecasting the presence and intensity of hostility on Instagram using linguistic and social features. In: *International AAAI Conference on Web and Social Media*.

Lo C (2018) *When All You Have is a Banhammer: The Social and Communicative Work of Volunteer Moderators*. Master's Thesis, Massachusetts Institute of Technology.

Mahar K, Zhang AX and Karger D (2018) Squadbox: A tool to combat email harassment using friendsourced moderation. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. pp. 586:1–13.

Massanari A (2017) #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society* 19(3): 329–346.

Matias JN (2016) The Civic Labor of Online Moderators. In: *Internet Politics and Policy Conference, Oxford, United Kingdom*.

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

Oldenburg R (1999) *The Great Good Place: Cafes, Coffee Shops, Bookstores, Bars, Hair Salons, and Other Gangouts at the Heart of a Community*. Da Capo Press.

Pater JA, Kim MK, Mynatt ED and Fiesler C (2016) Characterizations of online harassment: Comparing policies across social media platforms. In: *Proceedings of the 19th International Conference on Supporting Group Work*. ACM, pp. 369–374.

Roberts ST (2016) Commercial Content Moderation: Digital Laborers' Dirty Work. In: Noble S and B T (eds.) *The Intersectional Internet: Race, Sex, Class and Culture Online*. Peter Lang Digital Formations series, pp. 147–160.

Seering J, Kraut R and Dabbish L (2017) Shaping pro and anti-social behavior on Twitch through moderation and example-setting. In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, pp. 111–125.

Slovak P, Salen K, Ta S and Fitzpatrick G (2018) Mediating conflicts in minecraft: Empowering learning in online multiplayer games. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, pp. 595:1–13.

Sternberg J (2012) *Misbehavior in cyber places: The regulation of online conduct in virtual communities on the Internet*. Rowman & Littlefield.

Wohn DY, Freeman G and McLaughlin C (2018) Explaining viewers' emotional, instrumental, and financial support provision for live streamers. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, pp. 474:1–13.

Yang Q (2017) The role of design in creating machine-learning-enhanced user experience. In: *2017 AAAI Spring Symposium Series*. pp. 406–411.

Appendix A: Interview Protocol

The following script was used for interviews, with questions on relationship with platform employees added subsequent to interviews with Twitch moderators.

Introduction

1. How long have you been active on [the platform]?
2. How long have you been a moderator for [specific community]?
3. Do you moderate any other communities? For this interview, please focus your answers on [community].

Primary Questions

1. How would you describe the culture of this community?
2. How did you become a moderator in [community]? [If they started the group: How did you select moderators for your community? General characteristics, needs, experience?]
   (a) Did you know other moderators in the group before you became a moderator?
   (b) Did you have prior experience as a moderator?
   (c) Were you active in the community before becoming a moderator? If so, in what ways?
   (d) Did you provide support to the community such as design, technical, financial, or other ways?
   (e) Did you volunteer or were you asked to become a moderator?
3. How did you learn about how to be a group moderator?
   (a) Were you ever formally trained by someone on how to be a moderator in the group?
   (b) Did you receive any instructions?
   (c) Did any other moderators give you guidelines or advice?
   (d) Were you given examples of scenarios that might come up and instructions on how to handle them?
   (e) Did you get a chance to practice in any way before starting to moderate the community?
   (f) Did you learn anything about how to be a moderator from [site]'s tutorials or explanations of moderation resources?
4. Do members of your team have specific roles? [If single moderator, or they say they all moderators do the same things: Can you tell me about the different types of things you do in managing this community?]
5. In what types of situations with the group do you have to step in as a moderator?
6. What types of violations have you spent the most time dealing with in the past week or so?
7. *Double check with them* - "It sounds like you do x, y, and z. Is that everything, or is there anything else?"

8. What are some technical tools you use to make your job easier?
   (a) Do you use any sort of screening to filter new members?
   (b) Do you find these tools to be sufficient?
9. In the past month, what sort of things have you spent the most time discussing with other moderators of the group?
   (a) What platform(s) do you use to discuss these things?
10. Do regular members of your community ever help with moderation?
   (a) Do they ever criticize people who break rules or explain to them how to behave in the group?
   (b) Do they ever report content that they think violates the rules?
11. Do you ever warn users before bans or post removals?
   (a) When are these warnings customized, and when are they templates?
12. Do you ever post or send explanations of the rules after removing a post?
   (a) When are these explanations customized, and when are they templates?
13. Can you give me an example from the past week or two of how an offender has reacted to being punished?
   (a) Is this a typical reaction?
14. How have the rules in your community changed over time?
   (a) What is the process for changing rules like?
15. Have you ever interacted with [platform] employees regarding your group? If so, about what?
   (a) How frequently do you think platform admins review your community for compliance to site-wide content policies?
16. Can you describe to me a significant or memorable moderation experience that you've had in this group? It can be positive or negative.

Conclusion

1. Is there anything I didn't ask about or that I missed that you want to add about moderation in this environment?

Appendix B: Interviewee Characteristics

Interviewee	Community topic	Gender	Country
F1	Pets	F	USA
F2	Games	M	Mexico
F3	Academics	M	USA
F4	Games	M	Australia
F5	Entertainment	M	USA
F6	Memes	F	USA
F7	Academics	M	USA
F8	Memes	M	USA
F9	Niche Interests	F	USA
F10	Niche Interests	F	UK
F11	Academics	M	USA
F12	Niche Interests	F	USA
F13	Memes	M	USA
F14	Niche Interests	F	USA
F15	Memes	M	USA

Interviewee	Community topic	Gender	Country
R1	Technology	M	USA
R2	Support	M	USA
R3	Academics	M	France
R4	Niche Interests	M	UK
R5	Games	M	USA
R6	Memes	M	USA
R7	Sports	M	USA
R8	Sports	M	USA
R9	Memes	M	USA
R10	Memes	M	USA
R11	Support	M	USA
R12	Games	M	USA
R13	Academics	M	UK
R14	Support	M	UK
R15	Support	M	USA
R16	Academics	M	USA
R17	Cars	F	USA
R18	Pets	M	USA
R19	Memes	M	USA
R20	Sports	M	USA
R21	Academics	M	USA

Interviewee	Community topic	Gender	Country
T1	Classic games	F	USA
T2	Tabletop games	F	USA
T3	Variety gaming	M	USA
T4	Creative	M	USA
T5	FPS games	M	USA
T6	Difficult games	M	UK
T7	MOBA Games	F	UK
T8	FPS games	M	USA
T9	Variety gaming	M	USA
T10	MOBA Games	F	USA
T11	Variety gaming	M	USA
T12	Variety gaming	M	UK
T13	MOBA Games	M	USA
T14	MOBA Games	F	USA
T15	Variety gaming	M	France
T16	Variety gaming	M	Sweden
T17	Tabletop games	M	UK
T18	Variety gaming	M	Canada
T19	Variety gaming	F	Canada
T20	Variety gaming	F	USA

Appendix C: Code counts

1. BEING AND BECOMING A MODERATOR				
Step	**Theme**	**T**	**R**	**F**
Becoming a moderator	Friend, family member, or connection	4	2	13
	Recognized from other moderating experience	3	3	4
	Stand-out member of the community	15	14	11
	Availability at important times of day	2	2	5
	Volunteered or applied to become a moderator	0	10	6
Role differentiation	No different roles	0	0	8
	There is a head mod and/or hierarchy	2	3	11
Learning to be a moderator	Discussion or instructions	6	5	11
	Implicit understanding from being in community	7	2	8
	Learning by doing	0	12	13
Communication between moderators	Discussion about moderation decisions	7	12	10
	External platforms are used for communication	4	10	4
	Internal platforms are used for communication	0	14	13
	Off-topic or social conversations	0	3	3
	There is little or no communication	2	5	2
Development of a moderation philosophy	Valuing direct engagement	3	2	3
	Hands-off approach	0	4	2
	Maintaining a neutral stance	0	2	4
	Moderators as group “police”	0	3	4
Relationship with site administrators	Little or no engagement		19	14
	Work together to address problems		10	2

Figure 5. Steps and variants in Being and Becoming a Moderator process [κ = 0.89]. Code counts by Twitch (T), Reddit (R), and Facebook (F)

2. MODERATION TASKS, ACTIONS, AND RESPONSES				
Step	Theme	T	R	F
Routine tasks for monitoring/maintenance	Approving new members	0	0	13
	Contributing to the discussion	6	2	8
	Keeping the space "clean" or managing potential conflicts	4	3	8
Incidents	Disruptive behaviors	13	12	5
	General incivility	16	14	14
	Targeted attacks	16	11	10
Community Members' Responses	Critiquing offenders, explaining rules, defending community	13	6	9
	Flagging or reporting content	8	19	13
Moderators' Responses	Banning, timing out, or muting users, removing content	18	12	15
	Explaining to users why they were punished	2	20	10
	Use of tools beyond bans/timeouts for moderation	12	17	6
	Warning offenders	16	10	15
Offenders' reactions to warnings/timeouts/bans	Escalate behavior or resist	8	12	13
	No reaction	0	7	2
	Reform or apologize	8	12	9
	Seek clarification or request review	6	9	5

Figure 6. Steps and variants in Moderation Tasks, Actions, and Responses process [κ = 0.70]. Code counts by Twitch (T), Reddit (R), and Facebook (F)

3. RULES AND COMMUNITY DEVELOPMENT				
Step	Theme	T	R	F
Changes in internal dynamics	Community evolves and/or grows over time	8	10	7
	Issues or problems arise	2	7	12
	Temporary special situations	4	0	2
Process for changing rules	Community input	0	6	1
	Discussion among mods	3	12	8
	Executive decision	20	4	2
External influences	Site rules	0	9	3
Internal influences	Personal values	5	0	3

Figure 7. Steps and variants in Rules and Community Development process [κ = 0.85]. Code counts by Twitch (T), Reddit (R), and Facebook (F)