When All You Have is a Banhammer: The Social and Communicative Work of Volunteer Moderators

by

Claudia Lo

B.A., Swarthmore College (2016)

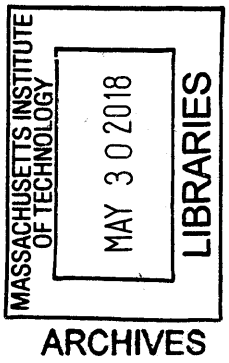
Submitted to the Department of Comparative Media Studies
in partial fulfillment of the requirements for the degree of

Master of Science in Comparative Media Studies

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2018



Signature redacted

Author .....
Department of Comparative Media Studies
May 11, 2018

Certified by ..... Signature redacted
T. L. Taylor
Professor of Comparative Media Studies
Thesis Supervisor

Accepted by ..... Signature redacted
Heather Hendershot
Professor of Comparative Media Studies, Director of Graduate Studies

When All You Have is a Banhammer: The Social and Communicative Work of Volunteer Moderators

by

Claudia Lo

Submitted to the Department of Comparative Media Studies
on May 11, 2018, in partial fulfillment of the
requirements for the degree of
Master of Science in Comparative Media Studies

Abstract

The popular understanding of moderation online is that moderation is inherently reactive, where moderators see and then react to content generated by users, typically by removing it; in order to understand the work already being performed by moderators, we need to expand our understanding of what that work entails. Drawing upon interviews, participant observation, and my own experiences as a volunteer community moderator on Reddit, I propose that a significant portion of work performed by volunteer moderators is social and communicative in nature. Even the chosen case studies of large-scale esports events on Twitch, where the most visible and intense tasks given to volunteer moderators consists of reacting and removing user-generated chat messages, exposes faults in the reactive model of moderation. A better appreciation of the full scope of moderation work will be vital in guiding future research, design, and development efforts in this field.

Thesis Supervisor: T. L. Taylor
Title: Professor of Comparative Media Studies

Acknowledgments

To T. L. Taylor, for her unwavering support for both my thesis-related and non-academic endeavours; to Tarleton Gillespie, my reader, for his generosity and thoughtful insight; to Kat Lo, fellow partner-in-academic-crime; to Shannon, CMS's very own chocolate-bearing problem-solving wizard extraordinaire. To my cohort, with whom I have endured this process and to whom I am indebted to for so much.

To the ESL moderation team who walked me through the baby steps of Twitch moderation with true grace; to DoctorWigglez, whose help left this thesis far richer.

To a certain verdant waterfowl, who taught me everything I know about moderation; to my moderation team on reddit (past, present and future) from whom I have learned so much; to the Euphoria regulars who provided me with feedback, support, and an uncanny ability to help me work out what I was saying better than I did myself; to the denizens of the Crate & Crowbar, menacing with spikes of pure wit and adorned with puns of the highest calibre, without which I would be short, amongst other things, a title; to the Cool Ghosts of the internet high-fiving me through the wee dark hours of the night as I made my way through the process.

To Erin, for everything:

My heartfelt thanks and deepest praise to you,
The seed of this was not of mine alone.
Without your constant guidance to turn to,
This thesis, stunted, would never have grown.
Yet with your care came blossoming of prose,
In ink it flowered and now lays in repose.

Contents

List of Figures

Chapter 1

Introduction

One night, as I was preparing to leave for class, I got a message notification from the chatroom that I help moderate. A user, posting in a specific area of the chatroom meant for LGBTQ users, asked if they were allowed to ask not-safe-for-work (NSFW) questions. This, in turn, sparked off pages of fast-moving speculation: what qualified as NSFW? How would the moderators respond? What policy had Discord, the platform provider, set out? In an attempt to assuage these fears, I ended up creating an on-the-fly preliminary policy regarding the posting of potential explicit content, while trying to deal with the reasonable concern that this LGBTQ chat would be cordoned off as an 18+ space, without going against Discord's Terms of Service. All while swiping wildly on my phone keyboard, trying not to fall down the stairs leading into the subway station.

The next day, my mod team and I evaluated the impact of this decision for both our Discord chatroom and our related community forum on Reddit. In the parlance of my fellow volunteer moderators, neither the banhammer nor the broom was needed: that is to say, no one needed to be banned, and no content needed to be swept away. Nothing was removed, so by the popular standards of online moderation, *no moderation had happened.* Yet this kind of decision-making forms the most significant and challenging aspect of my work as an online volunteer community moderator. Popular perceptions of online moderation work, both volunteer and commercial, portray it quite differently. Disproportionately, the discourse surrounding moderation, and related topics of online abuse, harassment, and trolling, centers on a small set of actions that I do as a moderator.

In my time online, as a participant and as a moderator on various platforms, the presence of human community moderators was common and everyday knowledge. Yet the discourse surrounding online moderation, particularly as it pertains to online social media platforms, takes quite a different route, instead hiding and suppressing the very presence of moderation as much as possible, and associating it less with human actors than to algorithmic processes and platform design.

The language of affordances here, drawing from Latour's actor–network theory, will be particularly helpful to figure out how the peculiarities of different online platforms shape the nature of the communities that they culture, as well as the forms of moderation and regulation that take place upon them. Online content moderation as it occurs on different online platforms has been the topic of increasing academic interest. I will borrow Gillespie (2018)'s definition of platforms for this thesis:

> For my purposes, platforms are: online sites and services that
>
> a) host, organize, and circulate users' shared content or social interactions for them,
>
> b) without having produced or commissioned (the bulk of) that content,
>
> c) built on an infrastructure, beneath that circulation of information, for processing data for customer service, advertising, and profit.

I am aware that 'platform' is a term that is thrown around quite liberally when discussing this subject. Hence, I want to distinguish between forms of moderation conducted *on* platforms, but *by* different stakeholders. The majority of work on moderation has centered moderation performed by platform operators, on the platforms that they run. Though it sounds tautological, this distinction is important: specifically I wish to divorce the idea that the reality of carrying out moderation on a platform always primarily rests on the platform operator. Such moderation work—that is, moderation performed by platform operators—has been variously described as intermediary governance (Gasser and Schulz, 2015), as the governors of online speech (Klonick, 2017), and as the self-defense of a semicommons (Grimmelmann, 2015). At this point, though, I would like to point out that such descriptions of

moderation render invisible the people who carry it out: those who design the boundaries of these places, who create the different tactics and policies that constitute this governing and regulatory work, and who ultimately carry them out.

Online moderation in all its forms has enormous impact on the experiences of millions, and potentially even more, as online social spaces proliferate. Yet even as current events make apparent the need for moderation in online spaces, we are, generally speaking, going into this practically blind. The public appetite for platforms to regulate users grows day by day, and yet we are unclear as to what it is we want, and how it should be done. Moreover, public discourse tends to place the responsibility of regulation upon platform operators alone; while for various political, rhetorical, practical and moral concerns, this may make sense, I fear that defining this argument with platform operators as the only group tasked with moderation blinds us to the efforts of community-led moderation.

I would propose two basic types of moderation: decontextualized moderation, and contextualized moderation. Decontextualized moderation is characterized by the fact that those who conduct this work are alienated from the community of users whom they are expected to moderate. Commercial content moderation as described by Roberts (2012) is one defining example: these moderators generally formally employed by the same company that runs the platform(s) upon which they moderate, but are distanced in multiple ways: geographically, being contracted workers far removed from the countries where the company itself may operate; technologically, by providing these workers with a controlled portal that does not allow them to seek out additional contextual information; and socially, by removing them from the platform itself. Additionally, the focus of commercial content moderation tends to be the content on the platform, rather than user behaviours or norms, and the actions that can be undertaken by these moderators is accordingly limited to either removing it, or leaving it alone. Decontextualized moderation would also extend to non-human agents that perform moderation work: the "algorithmic configurations" (Humphreys, 2013) that promote, suppress, and otherwise shape the contours of online social platforms. Examples of commercial content moderation might include an Amazon Mechanical Turk worker paid a few cents per image to decide whether or not an image is impermissible, either performing such moderation work directly or providing the human judgment necessary to train AI

other machine learning algorithms to eventually perform this work.

Contextualized moderation, on the other hand, is generally performed by people drawn from the same communities that they then moderate. This may be paid, as in community management positions, or unpaid volunteer labor. They work with a community that is expected to persist, forming expectations and norms that will impact moderation work. There are many striking similarities between professional and amateur contextualized moderation work. In brief, both "sit between developers and game players in the production network but are an important element in maintaining capital flows" (Kerr and Kelleher, 2015) although the presence or absence of an explicit contract will affect the relations of power as well as the responsibilities and implications of the role. Different platforms will additionally have different affordances which further impact the work of community moderators. The focus of their work is on the well-being, conduct, goals, and values of a given community, which encompasses any content created or found within it.

It should be noted that these are not mutually exclusive forms of moderation. Indeed, I would be surprised to see a platform that employed only one or the other regardless of their rhetoric. Community managers and moderators may employ forms of decontextualized moderation to do their work; for example, they may employ contracted workers in order to train automated tools, or use algorithmic methods to implement moderation policies that they devise. Conversely, the outcomes of decontextualized moderation may impact the work of contextual moderators; the rollout of a platform-wide moderation algorithm affects the work that the embedded moderators of that platform will then perform.

Additionally, these different types have different strengths. Most notably, contextualized moderation relies in some part on understanding the cultural and social norms and values of a community, thus presupposing the existence of a community in the first place. While the term itself is often thrown around by platform operators themselves, referring to 'a Twitter community' or 'a Facebook community' or 'a Reddit community', it is safe to say that such a platform-wide community exists only in the abstract. I turn to Preece (2000) for a working definition of community:

> An online community consists of:

1. People, who interact socially as they strive to satisfy their own needs or perform special roles, such as leading or moderating.
2. A shared purpose, such as an interest, need, information exchange, or service that provides a reason for the community.
3. Policies, in the form of tacit assumptions, rituals, protocols, rules, and laws that guide people's interactions.
4. Computer systems, to support and mediate social interaction and facilitate a sense of togetherness.

That is to say, these elements shape and direct moderation, and moreover that any given platform supports not one but a myriad sub-communities, with no guarantee that any one of their four constituent elements respect the boundaries of platforms on which they operate. As we continue on to look at moderation of communities, it is important to note that these elements are at once keenly felt by their members, yet also flexible, ambiguous, and fuzzy with respect to their borders. Thus, even as community moderators react to nature of said community as it pertains to their work, there is a degree of flexibility and deep cultural awareness at play.

There is a basic, formal distinction exists between different classes of online community members on the Internet: simply put, volunteer moderators are users given limited, but disproportionately powerful, permissions to affect what other users can or cannot see and do on the platform. This puts them uncomfortably between regular users, with no such special permissions, and platform operators or administrators, who have full permissions to affect the running of the platform; in the most extreme case this constitutes access to the literal on/off switch for the servers themselves. In a platform where some level of regulatory power is distributed, for example manipulating a comment 'score' that in turn affects the comment's discoverability, one would expect a moderator-user to have permissions above and beyond this base level. On Reddit, where every user has the ability to manipulate comment score through a single up- or down-vote, per comment, a moderator can remove it, rendering its score moot; this would be an example of that 'disproportionately powerful' ability to affect content for other users. However, moderators on Reddit cannot shut down

other communities, or ban users from the platform itself; those permissions are only granted to administrators, who are employees of Reddit. In that sense, a moderator both has editing permissions beyond that of a regular user, but below that of an administrator or platform operator.

These volunteer moderators have been variously portrayed as exploited by capital as part of the "free labor" that keeps the Internet running smoothly (Terranova, 2000), or analyzed through more of a co-creative lens as laid out by Jenkins (2008). However, this model of user-generated moderation, distinct from various forms of commercial content moderation (Roberts, 2012), has been complicated in recent years. Rather than understand these users merely as exploited users, or as equal creative partners, volunteer moderators work within an alternate social structure of values, motivations and norms that is influenced and shaped by capital and existing structures of power, yet does not necessarily respect their boundaries and strictures. This is a similar complication to that raised by Postigo (2016) in his 2016 analysis of YouTubers and how they generate revenue from the site.

Volunteer moderators and volunteer moderation has been described in many different ways, as peer-produced oligarchic institutions (Thomas and Round, 2016; Shaw and Hill, 2014), as sociotechnical systems (Niederer and van Dijck, 2010; Geiger and Ribes, 2010), autocratic "mini-fiefdoms" enabled by platform policy (Massanari, 2015), performers of civic labor (Matias, 2016), moral labor (Kou and Gui, 2017), and as negotiated peer regulation and surveillance (Kerr et al., 2011). The wide range of these descriptions suggests an equally broad subject matter: that is to say, moderation, in different spaces for different communities, may be called upon to perform all manner of roles. Nor do I believe any of these are necessarily mutually exclusive. Much like the mixed reality of contextualized, decontextualized, professional and amateur labor that comprises online moderation, what exactly moderation *is* is equally mixed and dynamic. Quite simply, the work of volunteer moderators, even a very narrow subset, is complex enough that we stand to benefit from a broader picture of that work, to better compliment what work exists. In particular, I want to locate the human workers in this work.

I am not necessarily proposing a 'how-to' guide for either platform operators or volunteer moderators. In contrast to broader work on online community building, such as Kraut and

Resnick's guide on building online community, I do not want to simplify this down to a matter of design. Rather, my aim is to refine our understanding of, and perspectives on, online volunteer moderation. If we think of moderation in terms of what platform operators are doing, what are we missing out? And if we think of moderation as the removal of content, what do we render invisible to ourselves merely by having such a narrow definition? It is valuable to fill out that part of the everyday work that makes online social spaces what they are, because

I will focus on a particular subset of contextualized moderation, event moderation. This is moderation of a community that is centered on a specific event, and is therefore limited in time and scope following the contours of that event. Event moderation may be conducted on multiple platforms simultaneously, and its audience come together because of the event and disperse once it ends, though they might rarely form the basis of a longer-term community. For very well established events that recur on a regular basis, such as a yearly esports tournament, a community of regular users may also develop, but generally speaking the majority of the audience are newcomers or brief visitors and therefore once may not expect them to develop the same kinds of interactions with moderators as more stable communities do.

More specifically, I work with at large-scale esports event moderators on Twitch. These moderators work for large esports tournaments, which might be expected to draw several hundred thousand concurrent viewers at their peak. These tournaments generally run for a few days, over a weekend, and are put on by organizations that work together with game developers to run tournaments. The games covered by the moderators interviewed included Valve's *Defence of the Ancients 2* (DOTA 2) and *Counter-Strike: Global Offensive* (CS:GO); Riot Games' *League of Legends* (LoL); Blizzard's *Hearthstone*; and most recently, Bluehole Studio's *PlayerUnknown's Battlegrounds* (PUBG). Of these games, the two most common were DOTA 2 and CS:GO, with both drawing huge crowds. Some moderators would attach themselves to a particular game, moderating based on their existing fan attachment to that game, but it was not uncommon for them to work for the same event organizer on multiple concurrent events, whether or not they featured the same game.

Why this narrow focus? My intention is to demonstrate that even for a population of moderators whose most common action is the removal of content, there are many more im-

portant aspects to what they do, and how they conduct it, that has been overlooked. Even allowing for the most generous fit for our current Furthermore, these overlooked aspects are not merely complimentary to moderation-by-removal, but integral in guiding their moderation, both in the judgment and in the labour of performing these tasks.

1.1 Methodology

I conducted nine in-depth interviews with Twitch esports moderators with ethical approval from MIT. All of the interviewees had worked on large-scale events, here defined as recurring semi-regular esports tournaments that were expected to draw concurrent viewer counts of over 100,000. The largest regular event that some of these moderators covered were the Majors for CS:GO, including 2017's record-breaking ELEAGUE Major, which peaked at over a million concurrent viewers.

I also sat in on two large moderator-only Discord servers, one for ESL moderators and another more general Twitch moderation server, under a marked "Researcher" account, with which I solicited these interviews. While I reached out to most of my interviewees, a few volunteered to be interviewed and would recommend others to me to be interviewed. Their quotes here have been lightly edited for grammar and to preserve anonymity. The questions for my interviews were largely based from some preliminary conversations I had with Twitch moderators, and also drew from my seven years' experience as a volunteer moderator on some large (over 100,000 subscriber) groups on Reddit. All interviewees were given a consent form to sign and were allowed to view an advanced copy of this thesis.

I was also granted moderator status on the `esl_csgo` channel, and engaged in participant observation, starting with the Intel Challenge tournament on 26 February 2018. This channel broadcast the Electronic Sports League's *Counter-Strike: Global Offensive* events, and would draw upwards of 90,000 concurrent viewers during live broadcasts. The moderation team were aware of my motivations for joining, and I was expected to perform the duties of a junior moderator in accordance with their guidelines and existing moderation precedent.

I used my own regular Twitch account for this, but was given access to the ESL moder-

ation guidelines and so changed my setup to fit. This meant that I had to set up two-factor authentication on Twitch using the Authy app, sit in on their moderator Discord to remain in contact through the event, and was granted access to Logviewer for that channel. I did not have access to any bot settings. My primary focus was moderating the newly-added Rooms, one for each team playing, and I was not focusing on the main stream chat. I also relied on the help of moderators to understand the different meanings of the various emotes and more famous memes circulating on Twitch. This is especially needed if trying to read a chat log, since many of the emote names are in-jokes that have since expanded out, and at any given time an emote may be used for its surface-level appearance or for the in-joke that it celebrates.

Twitch event chat, with its fast pace and emphasis on repetition of in-jokes and memes, can be extremely intimidating at first pass. However, many features of moderator-facing chat clients, available for free, are also immensely helpful for researchers. It is vital to note that, as Twitch undergoes constant updates and revisions to its APIs, these third-party tools are liable to break or suddenly have limited functionality until their developers can support whatever is the latest version of Twitch chat.

Generally speaking, many of the features that moderator-facing clients, or other third-party plugins, implement are also extremely useful for researchers trying to get a grasp of the size, scale, speed and tone of a given channel. For watching chat live, plugins that allow for pause-on-mouse-hover are invaluable for keeping up with chat, which is offered by FrankerFaceZ[1] or Better Twitch TV, though the former has far more active developer support.

When watching live, other moderator-facing clients I used were 3ventic's Mod Chat Client, and CBenni's MultiTwitch. Multi-Twitch allows one person to see several chats next to each other at once, while the Mod Chat Client highlights and states which messages have been removed, and crucially, by whom. However, neither of these programs generate logs, and therefore are useful only if the researcher is also taking notes during the stream.

To generate logs of Twitch chat, I used both an IRC client, Hexchat, and a custom Twitch chat client, Chatty, which was designed for moderators. Both create chat logs as .log files

[1] As of time of writing, FrankerFaceZ has most of its features disabled as a new version is written for compatibility with the latest updates to Twitch chat. It also cannot affect Twitch Rooms.

which can be easily converted into a plain text file. Both of these clients can log ban or timeout messages, and preserve messages which are later deleted. In a plain text format, emotes are not preserved; instead they are represented by the name of the emote. Chatty also allows one to view a moderation log and AutoMod settings if the account used to connect to the channel has moderator status. However, even without moderator status, Chatty is extremely useful as it allows for keyword or keyphrase highlighting, looking at individual users' chat history, and to see charts of viewer count over time.

Twitch chat logs were also downloaded after-the-fact using a Python script, rechat-dl. This script downloads the JSON file containing recorded chat information, and this was converted to a CSV file using R, while also stripping out extraneous information. It should be noted that Twitch currently allows viewers to leave timestamped chat messages on replayed streams, meaning that this should not be understood as a perfect archival copy of the stream and stream chat. The stored messages also have some formatting quirks which must be dealt with; one major one is that deleted lines are preserved only as blank lines. This means it is possible to see how much of chat was deleted, but it is impossible to guess why, or to see who performed this action. Another minor issue is that timestamps are saved in epoch time, and I am currently unsure what Twitch uses for their epoch date.

Chapter 2

Models of Moderation

2.1 The reactive model

I call the existing understanding of volunteer moderation work the reactive model. At its core, this model positions moderators and their work as perpetually reactive, responding to what users do. It is both a narrative of moderator action, and an ideal. The narrative is that a user does something, a moderator sees it, and then the moderator either decides to do something or nothing in response.

The ideal form of moderator action is seamless, in that it should leave no or minimal trace. For example, if a moderator removes a comment, the remaining trace should not draw attention to itself, or it should be totally invisible. This is because moderator action is seen as an exception to the normal user experience. If "nothing" is what moderators normally do, when moderators do something, it is imperative that the disruption be minimal.

By and large, moderator action is conceived of as the removal of user content, or of users themselves. There exists a sizeable taxonomy of different forms of moderator actions aimed at removing content, or otherwise putting up a barrier to its legibility. Aside from total comment deletion, different platforms offer moderators different tools: for example, the practice of "devowelling" or "disemvowelling", where all the vowels in a given comment are removed, was a popular way to make disruptive comments harder to read (Kraut et al., 2011). The different forms of banning users are equally diverse. There are bans based on duration (temporary versus permanent versus "kicking", which does not stop one from logging in

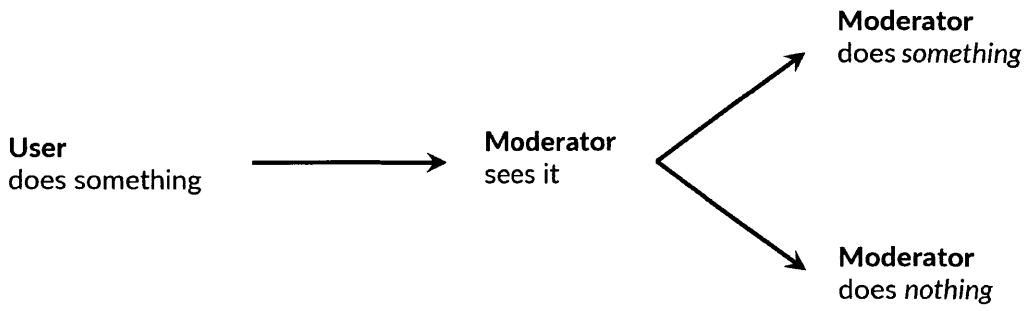


Figure 2-1: A diagram of the reactive model of moderation.

to the space again); there are bans based on identifiers (username bans versus the more extreme IP ban); and lastly there are bans based on formality (a regular ban that gives the banned user a notice, versus shadowbans, where any comment made by the banned user is immediately removed without that user's knowledge). Automated tools, whether third-party or built into the platforms themselves, further expedite this process by giving moderators access to blacklists of words, phrases, or more sophisticated pattern-matching tools. This automation also revolves largely around the removal of content.

Outside of removal, some platforms are beginning to give moderators ways to promote content, or otherwise positively interact with content. These may take the form of promoting, distinguishing, or otherwise highlighting comments or users whose behavior or content exemplifies values held by that particular community. For example, in Sarah Jeong's *The Internet of Garbage*, one suggested "norm-setting" moderation action is "positive reinforcement", the demonstration of good content or behaviours However, these positive interactions are still about reacting to content; a similar practice is suggested by Grimmelmann (2015). The only difference between such promotions and content removals is that, instead of obscuring content, positive moderator reactions make selected content easier to find. The tools available for positive moderator actions are also relatively unsophisticated; in contrast to the adoption of automated tools for comment or user removal, promotion of content is still largely done according to a moderator's human judgement. This reactive model explains

why total automation of moderation is seen as a plausible solution to online harassment and general negativity on a site. If moderation is seen as spotting red flags—words, phrases, emotes, usernames, avatars or similar—and responding to them, then total automation sacrifices little for a lot of potential gain. Under a reactive model, human judgement is only needed in the most ambiguous of cases, an outlying scenario easily handled by a skeleton crew. Proper punishment, having already been matched up to an appropriate behavior in internal moderation policy, can be doled out by the same tool. Automation would provide far greater reach and pace that human moderators could hope to achieve, at a fraction of the costs.

The overwhelming perception of moderation work is that removal of content lies at its core. This, therefore, carries with it connotations of censorship and punishment, as the perceived silencing of users runs counter to values of free speech and open expression that are so dear to online communities. Additionally, by positioning moderator action as an exception to normality, any visible moderation becomes a sign of emergency or crisis. The powers given to moderators are to be rarely exercised, and when they are, they carry with them enormous anxieties over the proper use of power.

This is not to say concerns over abuse of power are not justified, but that the fear of this happening is over-emphasized in the collective imagination of online communities. The reactive model's assumptions, that moderation work is punitive and exercises of moderating power herald a crisis, means that all such action—when visible—is scrutinized. The codification of moderating rules or guidelines, by which moderators are meant to act, become vital; objectivity becomes a virtue to which moderators should subscribe, mirroring popular understandings of criminal justice proceedings. Through this prism, moderators occupy a position of awful power. Unlike users, they possess administrator-like powers to remove content and remove users, with the additional ability to read this other layer of invisible, removed content; unlike administrators, they are not employed and therefore not clearly answerable to the same hierarchies, and are much less distant and more visibly active within the communities they govern. This uncomfortable liminality generates fear and anxiety, understandably, especially since moderation itself is meant to be invisible work. "Mod abuse" becomes a rallying cry against threats of moderator overstep, real or imagined. Therefore,

containing moderators by holding them to standards of transparency and impartiality, generally accompanied by the aforementioned formal written rules of the community, becomes imperative. Paradoxically, the need to have clear guidelines ahead of time is itself anxiety inducing as this is akin to an admission that moderation will be required, which itself is an admission of things going awry. Visible development of moderator policy, even without accompanying action, is meant to come only when those policies are tested or in other such emergencies.

The reactive model creates and sustains these contradictions, while at the same time obscuring important elements of the relationships between users, moderators, and administrators. While the basic definition of users, moderators and administrators still holds, the reactive model creates a strict boundary between users and moderators as a creative class of users versus a reactive class of users. It also creates a binary between users and non-users, since "moderator" in the reactive model is not generally clearly defined in the positive, and more broadly means anyone with more permissions than regular users. Even within its narrow scope, it is ill-suited to explaining or accounting for moderator considerations directly related to responding to content. For example, by tying all moderator work to direct reactions, short-term moderator action is disproportionately emphasized while longer-term moderation work tends to be overlooked. Distinctions between different moderation roles are also collapsed, as a consequence of the reactive model highlighting individual moderator actions as the key area of focus.

Additionally, blurring the distinction between different types of moderator leads to its own problems. Moderators are rarely recognized as a distinct group of users under the reactive model. They are either lumped in with users, or assumed to be acting in the interests of administrators. Thus, calls to expand volunteer moderation efforts are sometimes interpreted as calls to expand moderation powers to all users, as we can see with the classification of visibility systems (such as likes or upvote/downvote systems), comment reporting systems, or even recruiting users en masse as comment reviewers, as moderation systems. Such efforts do not involve a distinct moderation class; in fact they actively blur the distinction between moderators and users. The vocabulary used for these systems may not reflect the implementation across systems, either: on Facebook, reporting a post flags it for mod-

eration by Facebook's own systems rather than volunteers; on Reddit, reporting a post flags it for the volunteer moderators of that subcommunity.

On the other end of this extreme, moderation is sometimes held as the sole reserve of administrators, who rely on opaque regulatory systems such as algorithmic content promotion, automated content filters, and the like, often to cope with the sheer volume of content that they must sift through. All of these regulatory systems are open to manipulation by bad-faith actors. In the former case, their reliance on user input means they are vulnerable to organized disruption. In the latter, case, their effectiveness relies on the precise calculation by which these automated tools prohibit or promote content remaining hidden. As soon as their mechanisms are understood, they, too, can be manipulated by organized groups to promote specific forms of content, or to sidestep filters.

Because it forms the basis for our unspoken understanding of moderation work, the reactive model has guided efforts to design and create moderation tools, or platform design more generally. For example, popular platforms such as Reddit, YouTube and Twitch have by and large recognized the need for a distinct moderator class, to which community members can be promoted by community leaders (the subreddit founder, channel owner, or streamer, respectively). However, most platforms have been slow to adopt more sophisticated moderation tools beyond the basic ability to remove comments or ban users. Facebook Live, one of the more egregious examples, still has absolutely no affordance for moderators, and it is impossible for anyone to ban users or remove comments on a Facebook Live video. YouTube's lack of granularity when it comes to assigning moderator permissions means that channel owners face a difficult decision. Because editing and moderation permissions are set for the entire channel, rather than, say, for individual contentious videos, or limited to certain actions, they must either give volunteer moderators almost total power over their own channels, or choose not to take on volunteer moderators. This makes sense if the imagined environment for moderation is a crisis situation, where it would be more expedient to give an emergency moderator maximum permissions instead of having to hunt through more granular settings. However, if moderation is part and parcel of a channel's daily operation, granting such wide-ranging power becomes more of a liability.

Using the reactive model as our basis for understanding moderation severely limits our

ability to collaborate with moderators (if this is even recognized as an option) and to create tools to limit or combat online harassment. If we believe that all moderation is reaction, then all the tools we create and the questions we pose revolve around faster reactions, rather than seeing if there are other ways to pre-empt harassment and abuse. By focusing on the event of removal itself as the be-all and end-all of moderation, we ignore the importance of longer-term community care as we help repair the damage caused by harassment, and build more robust methods for dealing with abuse of all kinds. We also ignore the fact that, by positioning comment removal as the default solution to dealing with harassment, we also default to letting the harassers get away with it, as if abuse were a force of nature rather than a set of conscious choices made by other human beings. If we assume the correct course of action is to remove an abusive comment after it has been made, we already accept that what the harasser wants—for their abuse to be delivered to a public forum, or for it to be read by their target—will always have already happened.

Lastly, the reactive model ignores constraints on moderator agency, as well as the way in which nonhuman agents such as platform design influence moderator action. While it is true that users lack the formal power that moderators possess, they are still capable of using "soft", social influence or other forms of resistance against moderator actions. Likewise, moderators may take actions that run counter to the desires of some of their users, or counter to the best interests of the platform's administrators. Moderators may voluntarily constrain themselves in accordance with ethical principles, or out of concern for the public fallout of their actions. Even within the realm of reactive moderation, there is a complex web of relationships, formal and informal, with their attendant tensions, considerations of power, expected short- and long-term consequences, to factor in to every decision.

2.2 The lasting power of the reactive model

The popularity of the reactive model, as a way to conceive of the entire issue of moderation online, is undeniable. To be sure, reactive content regulation work is a significant part of what moderators do, and this has also shaped the ways in which moderators relate to, and understand, their work. Content regulation is also the most visible aspect of moderation

work from both user and administrator perspectives. There are also practical considerations that push moderators to hide themselves and the precise way their work functions from the users they regulate. For example, certain content removal systems, especially those that rely on relatively static, less-flexible nonhuman filters, need to be kept opaque in order to be effective. Additionally, visible traces of moderation may hail crisis, but are themselves a visible scar that disrupts the experience of other users. New technologies that hide these traces (for example, removing even the "message deleted" notification) do provide a better user experience in that it further minimizes the damage caused by the content that required regulation. To put it another way, the crises and anxieties that visible moderation dredges up with it are not necessarily unfounded.

The reactive model also simplifies the problems of moderation into a neater form, which is more solvable. By portraying moderation work as a simple chain of cause-and-effect, and focusing solely on those areas of moderation work that is the most conducive to this portrayal—content regulation—the complex messiness of moderation work is cleaned up and becomes a problem that can have a solution. When moderation work is no longer about relatively fluid groups of people acting on, with, against and for one another, with multiple motivations, abilities, valences and outcomes, it itself becomes something that is manageable, categorizable, and controllable. Pared down, it is easier to operationalize and automate. A study conducted by Kou and Gui (2017), of players participating in Riot Games' Tribunal system, points out that for all its lauded successes, the Tribunal was quietly replaced by an automated solution, despite the fact that "players repeatedly questioned the automated system, citing its opaqueness, vulnerability, and inability to understand human behavior." To this day, the Tribunal system is still offline. Granted, we do not have information on the efficiency of the Tribunal system later on, nor of its automated replacement, but the social work that the Tribunal system performed—allowing participants to engage in moral labor, whether or not this occurred because Riot "[convinced] people that it was righteous to participate" (Kou and Gui, 2017)—does not seem so highly valued.

Acknowledging the social and affective dimensions of moderation, by contrast, means acknowledging that human judgement is replaceable or reducible in the work, and that there is a human toll on the workers who perform this labor. To frame moderation work as an

exercise of punitive power requiring objectivity and rationality is to place it within a hierarchical system, with users at the bottom, moderators in the middle, and platform operators on top. I believe it would be a serious misstep to leave all oversight to corporations, or even position them as the ultimate arbiters of behavioral regulation. Moderators and users, as we shall see, have a robust understanding of and ways to deal with abuses of power even in the absence of official tools or formal support for these deliberations. They are capable of dealing with problems of power abuse and policy changes in a way that is attentive and responsive to their needs, desires and values.

2.3 Counter-model: the proactive model

The proactive model grows primarily out of my experiences in community moderation on Reddit. Though the particulars of moderating on different platforms are of course distinct, the underlying ethos and domains of moderation work remain relatively consistent. The proactive model is my attempt to expand the reactive model. The work of reacting to content does consist a significant portion of moderation work, and certainly moderators' own understandings of what they do is deeply influenced by this reactive work. Yet, this is not the sum total of moderation work. While we may believe that this work is, or ought to be, practically focused, objective, rational, and ultimately about the application of regulatory functions by human or nonhuman actors, the day-to-day reality is more complex. Moderators frequently engage in social and communicative work, coordinating between users, fellow moderators, broadcasters or other personalities within their communities, and platform administrators. They also engage in civic labor where they create and amend policies, but also respond to policies or policy changes set forth by platform operators (Matias, 2016). Their work is vital in creating a cohesive community (Silva et al., 2009) through the use of soft social skills, not merely through the removal of un-permitted content.

This alternative model is less a replacement of the reactive model, and more an overhaul of the same. Rather than think of moderation as comprised solely of discrete events where moderators exercise regulatory power over the users they govern, the proactive model places these exercises within a wider trajectory and backdrop of moderation. This trajectory

is formed from the interconnected practices, norms, behaviours, values, and affordances found in the technical landscape in which both moderators and their community are situated, the social field of moderation (both specific to that platform, and broader cultural considerations of moderation), and the reflexive social, mental and emotional work that volunteer moderators conduct in order to comport themselves as moderators.

Furthermore, the temporality of moderation must be expanded to encompass the preparatory work of moderators, and the ongoing social processes that make meaning out of regulatory work and fold it back into the service of changing or bolstering pre-existing social attitudes and values regarding good or proper moderation. Moderators accumulate and preserve knowledge gained from previous experiences, and are in constant dialogue with their collective memory of these past exercises as it both guides and pressures current moderation actions. Formal or otherwise, they remember and collect information on different actors and use those memories in order to navigate moderation work. However, it is also important to note that these are interpretive exercises: the relationship between past precedent and current action is negotiated and interactive, mediated by the communicative and archival affordances of whatever platforms and tools moderators can access, as well as the values and norms of the moderator community in question.

Additionally, it recognizes that regulatory behaviour is not the only kind of work that moderators perform. Under the proactive model, the technical work of developing, maintaining, and adapting both in-built and third-party tools for moderation would qualify as "moderation work", as would emotional and mental health work conducted by moderators for their communities and for each other. Lastly, it complicates the position of moderators within their existing networks by acknowledging the impact of other groups, such as users, external organizations or corporations, platform operators, and other relevant parties, on moderators and their behaviour.

This more holistic understanding of the work of volunteer moderators uncovers their invisible work, in order to better appreciate the work already performed by these actors. Without a broader understanding of volunteer moderation work, efforts to improve social spaces online may well fall short, as we neglect a key group that already has a robust history of employing, creating and adapting whatever resources are available to them in order to

perform this kind of community-building work. With a better knowledge of what it is that moderators do, we can better create the infrastructure and resources necessary to support them in performing this crucial labor. The invisibility of moderation work need not remain unacknowledged, even as it remains largely unseen.

Chapter 3

Moderation on Twitch

Twitch.tv is an immensely popular livestreaming site, which allows its users to host live video of themselves. Launched in 2011, Twitch has its roots in Justin.tv's games livestreaming section, before it became big enough to split off. Twitch is easily the dominant livestreaming platform for esports and other major game tournaments, although recent efforts by other companies such as Google (through its YouTube Gaming program) and Facebook have emerged as alternatives.

Twitch is split up into different channels, each controlled by a single streaming account. Viewers largely interact with streamers through a live chat panel that can be found to the right of every single stream. As viewers on Twitch settle on particular channels that they deem their favourites, communities form. These same communities may come together to view one-off events, exchanging norms, ideas, and on the most superficial level, different memes, emotes, and new ways to spam messages in chat.

Twitch chat is the site of complex interaction between streamer, chat user, and moderator. Through repetition and iteration of simple, easily copied-and-pasted messages (commonly called "spam" in the parlance of Twitch), the effect of a crowd of fans roaring for their favourite teams is replicated. However, spam is rarely so straightforward: in-jokes abound, and remixes of existing spam to change the meaning are common. Moderators on Twitch ostensibly are primarily charged with managing this often unruly crowd, regulating the types of speech found in chat. [1]

[1] to be expanded

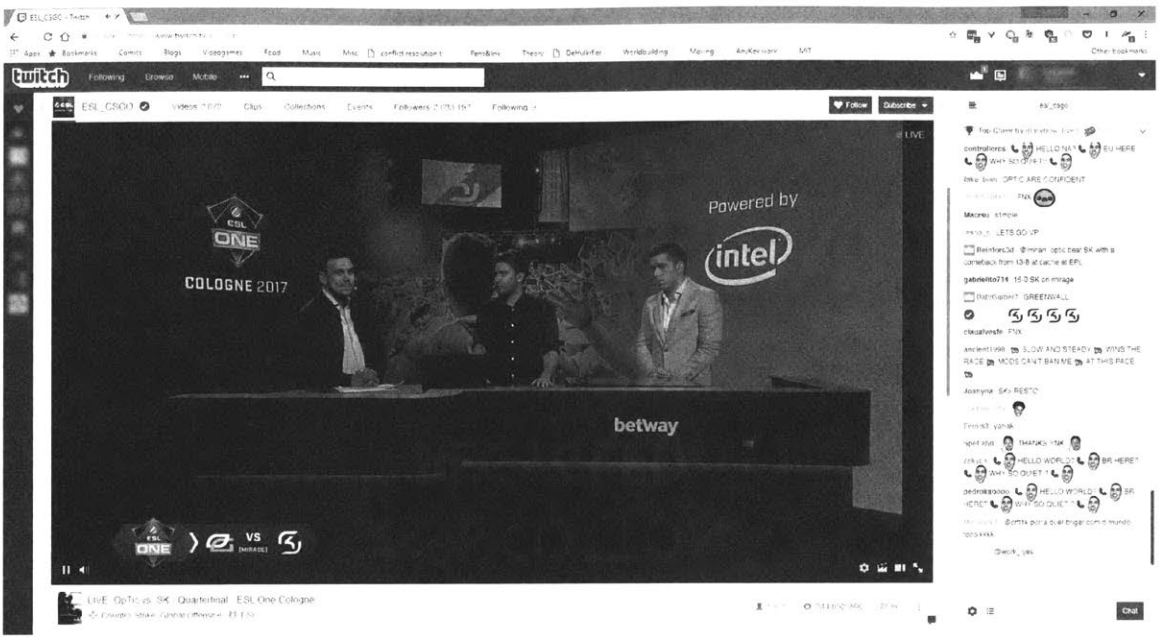


Figure 3-1: A screenshot of Twitch.

The two most commonly seen types of volunteer moderation on Twitch are community and event moderation. It is important to note that these forms of volunteer moderation are not diametrically opposed, and oftentimes the same person will take on both roles on an as-needed basis.

Community moderators are volunteer moderators that typically take care of a single streamer, or network of like-minded streamers. In addition to regulatory work, they may take on additional community management work. For example, a community moderator may take it upon themselves to greet newcomers to the channel, or to redirect viewers to different resources to learn more about the stream. This might expand out to managing auxiliary fansites, or cultivating different specializations: a moderator might be responsible for producing custom graphics, or managing different bots, tools, or scripts used in the channel. Because the purpose of establishing a channel on Twitch is to cultivate a stable repeat audience, and to steadily grow it, moderators end up forming relationships with the viewers, especially channel regulars. The flow of chat is also more of a back-and-forth between the streamer and their viewers, producing a more conversational atmosphere. The lengthier broadcast times might also lead to more hours worked per moderator, on a more consistent basis than event moderators.

As previously mentioned, event moderation is centered around discrete events that typically last no longer than a single weekend. Chat during event livestreams are rarely conversational, mostly consisting of audience reactions to things shown on-screen, or repeated fan-chants, memes and spam, due to the sheer volume of comments. Events draw high viewer counts, but these viewers are unlikely to stick around and form any basis for a permanent community. Instead, they coalesce for each event. While there is some overlap between different community streams and events (for example, the regular viewers of a esports professional may be likely to chat during events where that professional is participating), events are rarely dominated by a single community.

3.1 Moderation tools

In order to expedite these kinds of mass moderation work, moderators turn to different tools, many of which are developed by third parties and far exceed what the platform itself offers in terms of moderation functionality. To generalize, moderators rely on tools modify Twitch's user interface, add in chat history or archival search functions, and create flexible automated chat filters that can be adapted on-the-fly as new situations arise. Moderators also use other applications and programs, for communication, chat monitoring, and other peripheral considerations. They also employ built-in tools and settings available in Twitch itself. Third-party tools do not wholly supplant Twitch's built-in tools, but greatly expand what they are capable of accomplishing on the platform, and some were considered by my interviewees to be indispensable to their work. These tools, however, are not officially supported by Twitch; they exist in a kind of careful dance around the official updates and rhythms of the platform, and every update is scrutinized by the maintainers of these tools as they represent new points of failure for their tools. This is the case even for updates that promise to help moderators by building on Twitch's inbuilt moderation options.

As Twitch's moderation tools developed, so have these third-party tools correspondingly grown more complex and powerful. Similarly, tools that are widely adopted by the moderation community and championed by respected individuals gain prominence and popularity. These tools should be understood as created and informed by the practices of moderators

paired with the relatively new ability for individuals or small teams of developers to use platform APIs, such as Twitch, to make tools that tap into those systems directly. Twitch's open (or more open) API afforded the growth of a moderation tool ecosystem, which in turn supported the growth and sustenance of a class of moderators. While an open API is not a necessary condition for a healthy moderation community to form, it is an important factor.

On extremely powerful tool moderators have at their disposal are bots. These are programs that automate many of the more common actions that moderators would be expected to perform; on Twitch, there exist many different bots aimed at different groups of users, with the majority aimed at helping out streamers. Bots that prioritize the needs of moderators are rarer, though many streamer-facing bots have features that make them suitable for moderation as well. Common functions that these bots perform include removing messages and users, permanently or temporarily, according to certain criteria, or dispensing information through the use of custom chat commands. Bots are often named after their creators, and are funded through many different models: some are wholly free, while others might have premium features locked behind a one-time or recurring payments. Of the moderators I interviewed, the two bots most commonly named were moobot and ohbot.

Moobot has a well-maintained graphical interface, allowing moderators to change its settings via an external dashboard. It can be programmed with custom commands, for example allowing users to message moobot in order to get schedule information for the event they are watching. Moobot also has an automated spam filter that stops messages with excess capitalization, punctuation, repetition, certain memes, as well as allowing a custom blacklist of phrases or words. It is donation-funded; while no donation is necessary to use its basic features, donating money to moobot gives a user points, with which they can unlock more features, and more slots for editors. Editors are users who have access to moobot's settings; for a moderation team, this might include anyone trusted with changing blacklist phrases or adjusting its filters.

Ohbot, by contrast, is far more difficult to set up. It has no user interface, meaning that moderators can only change its settings by typing commands into chat. However, it is one of the very rare bots that is meant primarily for moderation. Its primary function is as a chat filter, and it not only allows for extreme granularity in settings but also can match

strings using regular expressions, or regex. Regular expressions are search patterns which allow for far more powerful search and pattern-matching capabilities than normal word or phrase blacklists. Using a standard syntax, regular expression strings can catch many variations on the same word or phrase, which means that a single well-tested regex string can have the same effect as multiple blacklist entries. Regular expressions can also be used to check the context in which a phrase is used, since it can check ahead or behind the phrase in question. For example, a regex string could be set up to permit 'TriHard', a global Twich chat emote depicting a black man. However, that same string could be set to match if 'TriHard' was embedded within a longer, racist message. Equally, a mispelled, poorly-tested or poorly-thought out regular expression string also has the potential to cause trouble. A misconfigured regex string might end up matching not enough or no messages, rendering it useless as a chat filter. Or, it could match too much, and incorrectly ban or timeout users sending innocuous messages, forcing moderators to reverse these bans, mollify chat, and take down the regex filter to be fixed.

Mastery of regular expressions is a specialist skill that is not common in all moderating circles, and those who understand regex are sought out by head moderators. Moderators who are responsible for creating these regex strings are guard them closely; one head moderator said that their regex string was seen by "about 6 people", all of whom had access only because they were also active contributors. One particularly well-known moderator's regex settings are now part of ohbot's presets, and the presence of a name helps prove its efficacy by explicitly giving that preset a respected author. However, not all moderators seek to learn it, simply because it is quite complicated, takes time and effort to learn, and comes with a different set of responsibilities. In the words of one moderator, "I'm good at what I do now, I don't want to to much more than that. I don't want to pick up everything, to do all the botwork and the programming behind that. I'll pass on that one."

Many other bots exist, with slightly different sets of features. The moderators I interviewed seemed to choose which bots they used based on their own familiarity with them, their moderation team's familiarity with them, the features available to their chosen bot, and whether or not the bot was in active development. Some of the other bots mentioned by my interviewees were xanbot, hnlbot, or Nightbot. One moderator I talked to also worked to

develop their own bot, to allow them to check accounts by age.

There is another bot available to all moderators, AutoMod, which is built into Twitch. AutoMod holds messages for human review, and uses machine learning to determine which messages should be held. AutoMod has four settings, which increase its sensitivity and change the types of messages that it targets; on the lowest, least-sensitive setting, it might only filter out racist slurs, while on the strictest it will remove all forms of hate speech, violent language, and any profanity. Though the moderators I spoke to appreciated it, especially once its had matured a little past its debut performance, they did not regard it as a one-size-fits-all solution to chat filtering. AutoMod can catch most general use forms of impermissible speech, but is relatively easy to circumvent by using emotes, memes, racist stereotypes or scene-specific in-jokes to express the same offensive sentiments. Moderators cannot respond by setting AutoMod to a higher setting, for two reasons; firstly, only the broadcasting channel account can change AutoMod settings; secondly, since the settings are relatively opaque and come in bundles, setting AutoMod to be more restrictive risks chilling chat to a degree deemed unacceptable by the moderators I spoke to. In my own observations, though AutoMod did filter out many messages, there were a significant portion of messages that were removed either by moobot, ohbot, or direct human intervention.

The second most common tool mentioned by my interviewees was Logviewer. This is a quasi-bot tool, which sits in channels that have opted in and generates a log of all messages that have been said in it. Moderators then log into an external site with their Twitch accounts, granting them access to the full chat history of the channel. Crucially, Logviewer allows moderators to see an individual user's chat history within that channel for as long as it has opted into Logviewer. Moderators can also add comments on a user, which can be viewed by all the other moderators.

According to my interviewees, Logviewer is useful because it allows moderators to keep a record of events. This is most useful when handling unban requests, or trying to sort out disputes between users. Since it also records the name of the moderator who performed bans, timeouts or other actions, the team ccan also determine who should be a part of any decisions for unbanning users. It is so useful and popular that it has been integrated with another popular third-party tool, FrankerFaceZ, and there exist other, even more special-

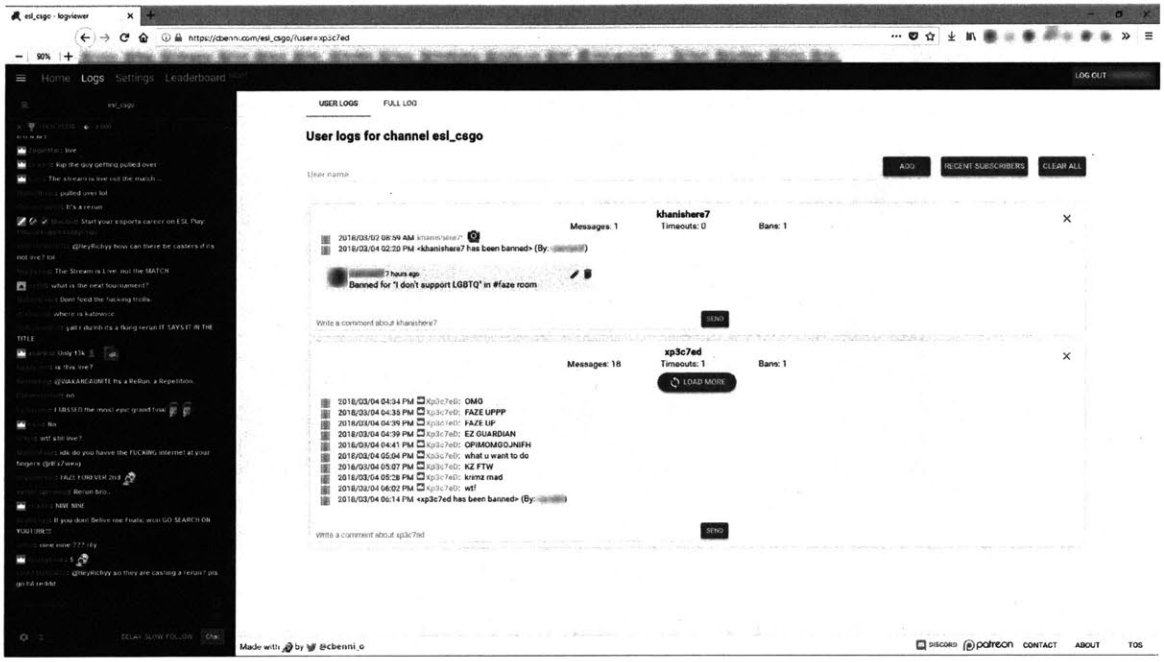


Figure 3-2: An example of Logviewer, showing multiple user chat histories, with moderator comments on a user.

ized bots that link Logviewer to other applications such as Discord. A similar logging tool, Modlog, exists specifically to track moderator actions. It allows head moderators to see who has been working and when, as well as to spot suspicious activity that might indicate a compromised account. Modlog similarly has bots that link it to Discord, so that moderators who are not present in Twitch chat are still notified of important moderator actions that a given team might want to review.

The last tool mentioned by all my interviewees was some sort of user interface improvement, and the two that were named were Better Twitch TV and FrankerFaceZ, or FFZ. Both are browser plugins, and essentially give users the ability to customize the way Twitch chat looks. Though neither was developed primarily for moderators, they are nonetheless widely used for this purpose. Of the two, FrankerFaceZ was preferred by the moderators I interviewed because of its more active developer, and because it performed all of Better Twitch TV's functions as well. FFZ's developer has also collaborated with the maker of Logviewer, to integrate Logviewer into FFZ, making it even more helpful for moderators.

These extensions primarily make it easier for moderators to perform common actions, such as timeouts and bans, and to view more information in one space. Figure 3-3 shows a

Figure 3-3: FrankerFaceZ's moderation card.

'moderation card', which FFZ brings up if a moderator clicks on a user's name. The moderation 'card' can display the user's chat history, and allows the moderator to choose from a series of timeouts (from five minutes to one week) in addition to adding a drop-down menu with a list of customizable ban reasons. It also adds hotkeys for timeouts, bans and purges, and can highlight messages that contain a particular phrase in chat. This, in conjunction with pause-on-mouse-hover, allows moderators to perform moderation actions at a much faster rate, and to keep up with the pace of Twitch chat..

Because chat moves by so fast, FFZ's user interface changes shine for moderators in large-scale chats. One of the most beloved features it adds is the ability to pause chat on mouseover, with one moderator calling it "a godsend," adding, "we also mod on YouTube, and I literally didn't mod YouTube for the past year and a half because you couldn't slow down chat, and things would just fly by and unless I could scroll up as fast as the chat was going it was just impossible." Issuing a timeout or ban involves either typing in the appropriate command in chat—`/timeout username [duration]` or `/ban username`—or clicking one of two small icons by a user's name. Without the ability to pause on chat, misclicks are very common and necessitate fixing, causing more work for moderators.

Lastly, in order to communicate with each other, moderators set up moderator-only groups where they can discuss moderation policy with each other. Since Twitch is set up to revolve around a video stream with accompanying live chat, it is not ideal for private group community discussions about individual moderation decisions. Instead, common ancillary

platformed used by the moderators I interviewed were Slack, Discord, and Skype. Skype has fallen out of favor with the advent of Discord and Slack, with moderators citing security concerns as a reason why they tend to shun organizing on Skype. Discord and Slack also have support for markdown, image and file sharing, and multiple text channels. These features allow for rich text formatting and easy sharing of screenshots or other files which make communication easier. Discord seems more popular for its gamer-oriented branding, granularity of permissions (through a role system, which also makes it a little easier to set up a moderator hierarchy), ability to set up voice chat, and support for bots using its API.

Within this third-party tool ecosystem, there are other artifacts that are not designed to directly help moderators work. Over my time observing this community, I saw a slew of other peripheral technical solutions, created to prop up this ecosystem. These included fixes, workarounds, add-ons, and other such kludgey solutions. These are not necessarily solutions meant to last, as the development cycle of both Twitch itself and these third-party tools rapidly forces them into obsolescence. However, instead of dismissing them as temporary elements meant to fade away over time, it is more useful to understand them as doing *bridging work* (Braun, 2013). These systems should also be seen and understood as part of the moderating tool ecosystem: involving the same actors, along the same networks, with impacts that may outlast the time in which they are in common use. The users who create tools for long-term maintenance may well be the same as those who push out quick fixes to make them work with the latest versions of Twitch, and add-on functionality may some day be incorporated into the tools that they enhance. Some examples of these bridging applications include tools that link Discord with Twitch, so that moderators can be working on Twitch, or monitoring it, without ever actually opening the site itself, and a Twitch Legacy Chat application, rolling Twitch's chat back to an older version supported by popular tools.

Aside from these plugins, applications, bots and scripts, there are a smattering of others that do not have a clear place within this complex constellation of tools. For example, Twitch requires the use of a third-party application, Authy, in order to set up two-factor authentication on one's account. Taking such security measures was something all the interviewed mods did, and so they all had to use Authy. Other useful moderating tools include Multi-Twitch, which allows a user to display multiple chat windows as well as multiple video

streams side-by-side in a single window.[2] Custom chat clients, such as 3ventic's Chat Client (designed for moderation) and Chatty, are also sometimes used, though they are not quite as flexible as FFZ. Moderators may also have mass unmodding or mass unbanning scripts on hand in case of compromised moderator accounts. However, my interviewes indicated that with the introduction of better security features such as two-factor authentication for accounts, the need for these drastic measure has lessened.

The value of these tools is most evident when they are absent. During IEM Katowice, for which I was a junior moderator, the moderation team decided to try and use the new Rooms feature of Twitch. Introduced in February 2018, rooms are side-channels, and allow a stream to subdivide its chat between a main chat and smaller peripheral rooms.[3] For this event, two rooms were set up, one for each team playing. However, this update had also broken nearly all of the moderation tools I described above. Only FrankerFaceZ had limited functionality, and even then, none of the moderation features were working. To get around this, the moderators installed another plugin which forced Twitch to use 'legacy chat', an older version of the chatrooms, which would be compatible with their tool suites. This had the side effect of making the new rooms invisible to both moderators and most of the moderator bots. Over the last day of the tournament, I spent most of my time sans tools, working only with the built-in moderator functions that Twitch provided.

Although the rooms were slower-paced than the main channel, it was still difficult to watch and regulate. If a single user was responsible for spamming the channel and making it impossible for others to participate, I then had to type in the appropriate timeout command as quickly as possible, with an appropriate duration in seconds. At a very basic level, this required parsing chat at least as quickly as it flowed, very fast and accurate typing, and memorizing different minute durations in seconds and making a judgement call to match the behaviour to a timeout length. If multiple users were participating in unpermitted behavior—for example, if one user encouraged many others to raid the opposing team's room by spamming insults to make it unusable—I had to call in help on the moderator-only

[2] See figure 3-4b.

[3] Interestingly, in the blog post announcing this new feature, Twitch suggested that moderators could use it as a private communication space. However, none of the moderator teams I spoke to had even considered doing so.

Discord room. Without the ability to look up user chat histories in the rooms, the moderators had to rely purely on their memory of events when adjudicating unbans after the event, generally erring on the side of caution.

Understandably, the presence of these tools was greatly appreciated by the moderators I interviewed. The ubiquity of these tools, and the streamlining they provide for moderation work, is so drastic that I would say that there is a generational gap between newer moderators who are used to these automated tools, and moderators who learned how to moderate in the absence of these tools. This sudden reliance on scripted tools, and the subsequent valuing of these new tools, did not seem to lead to jealousy or resentment in the moderator circles I studied. Instead, the prevailing atmosphere seemed to be one of gratitude, recognizing how much easier these tools had made online moderation. One self-described "old fart" moderator described this as "They [newer moderators] incorporate what they are studying, for example, into moderation...I see these young guys and I try to include them within the community more, and suddenly they rise way above me, for example in title, in certain areas. That's just amazing to watch, because not only did they make the entire community better, but they're getting the respect that they deserve."

It is important to remember that the explosion of moderation tools for Twitch did not arise *ex nihilo*. In the words of one moderator, "what the platform lacks is always that which the community creates itself." A common theme was that all the tools available to moderators were developed out of necessity, to fill gaps in Twitch's built-in moderation tools. Early on, this was done in the absence of these tools; however, the creation of more sophisticated functions such as Twitch's AutoMod has not diminished their importance in any way. In fact, one moderator stated that they were working on a bot that would have performed a similar function.

> I was teamed up with a friend to make a bot. The point of the bot was to analyze messages in channels, and determine what sorts of messages result in a ban, using machine learning [to] begin estimating what messages we should ban, and if that becomes solid enough, release that model and start banning messages using the bot, or warning mods that those messages should be banned. Literally a month after we started working on this project, Twitch released AutoMod,

which essentially does the same thing.

The development of these moderation tools is made possible by the affordances of the platform, the ease by which moderators can learn to create them, and the practical experience and knowledge generated by moderators guiding the creation of moderation tools. It is clear that moderators are not bound by what moderation options are built into the platform; rather, they are more restricted by how open said platform is. A few of my interviewees said that they disliked moderating on Youtube's livestreams, partly because of moderation input lag, but also because the moderation ecosystem was tiny in comparison to Twitch's.[4]

I interviewed two moderators who created and maintained third-party tools, and both of them mentioned that they had created these tools for themselves, and realized the need for these functions once large numbers of moderators asked them for access. When asked if crowdfunding platforms like Patreon might affect the mod-tool landscape in the future, both of them were skeptical, pointing out that even the most successful moderation tools did not bring in any appreciable profit for their creators, and even the most successful moderator Patreons would just about cover server and operating costs. With little monetary motivation to create and sustain these tools, it seems that the primary driver behind moderation tool development is the need and desire for better moderation tools, created by moderators for this relatively small group.

3.2 Running an event

Moderators do not simply show up at the start of an event and begin to remove messages; in an ideal situation, they will have prepared for this days in advance. Each event should also be understood as an opportunity for new moderators to learn how to moderate, for more experienced moderators to refine their skills and style, and for those moderators implementing new tools to test out their creations. Each event also has its own specific requirements and regulations coming from event organizers and the makeup of the presumed (and actual) au-

[4]Though Youtube's live streaming API has been available for many years, the moderators I spoke to only mentioned one moderation bot of note, Nightbot. Nightbot itself started off as one of the first Twitch moderation bots.

dience, necessitating on-the-fly adjustments by moderators. What follows is a generalized and ideal account of the steps involved in moderating a large-scale esports event on Twitch.

3.2.1 Preparation

In the run-up to an event, the head moderator may be contacted by a company liaison with relevant information. The head moderators I talked to said that this liaison usually was the social media manager, but ideally would be anyone with access to the broadcasting account on Twitch and contact with the production team. Head moderators pass on the account names of those who have volunteered to act as moderators, since the broadcasting account must be the one to make them moderators for the channel. The company liaison, in turn, gives head moderators information about the tournament schedule, predicted match start times, the teams in the brackets, the names and Twitter accounts of the casters, commentators and analysts, and other miscellaneous information such as the names of the songs that will be played during the event, or any promotional events such as giveaways that will happen during the tournament.

The moderators then work to set up the bots for the event; for the moderators I interviewed, this consisted of setting up moobot and ohbot. Moobot is a donation-funded bot that has a graphical user interface, making it much easier to use. It is capable of setting filters on chat, such as punctuation filters which stop messages that have over a certain ratio of punctuation-to-message-length, capitalization filters, and an easily-changed blacklist for phrases or words. It can also be programmed with custom chat commands, which can be triggered either by moderators or users. The head moderators, or any moderator with access to moobot ahead of the event, will wipe previous blacklists or custom commands so as not to trigger false positives when the event starts. New custom commands for displaying the schedule, score for the match being streamed, song titles, analyst information and more will be added using the information received from the company liaison. These generally take the form of `!command`, which users can use in chat to get that information. For example, a viewer who wants to know the day's schedule could type `!schedule` in order to get a message from Moobot that would link them to a webpage with the day's schedule.

At the same time, head moderators look for people to help work the event. Moderators

might be approached individually on recommendation, or recruitment calls might be posted in moderator community spaces. The recruitment process is fairly informal; the measure of a new moderator seems to be in the amount of work they do and their willingness to remain an active moderator, with a moderator's reputation in the community serving as a major factor in whether or not they find more moderation work. Newly recruited moderators are brought into event-specific moderator chat groups, and are told what this particular event's guidelines for chat are. Not every organization or developer will be the ones to issue such guidelines, and one head moderator said that "most people just leave it up to us [moderators], that have been doing it the longest."

There is no set consensus on how large a moderation team should be for an event of any given size. In fact, some moderators say that a few—numbers range from two to five—very active moderators are sufficient even for events of over 100,000 viewers. "The ELEAGUE [*CS:GO* tournament] was the biggest thing ever, a million people watching. And it was probably four of us active. There were a lot more people there, but only four of us were active." Other moderators concurred, and the phenomenon of a few active moderators managing an entire chat kept resurfacing. "there are two or three moderators who do 70% of the stuff. It's always like this. Most of the time it's the same people, the same very active people...and maybe five more who are semi-active, and maybe ten or fifteen others who do 1% each."

However, head moderators do try and recruit moderators with different availabilities in order to cover as much of the event as possible. Head moderators often try to recruit people from different timezones, which can be difficult for certain timezones when looking for English-speaking moderators.

> It's a typical thing, that it's very hard to find people from the Oceania or Eastern Asia regions...There's so few people that know English well enough to do moderation, there's so few people from that region in the community itself, it's basically just people from Australia. The population is so low that—it is ridiculously hard to find people from there!

Equally, a moderator might prepare themselves for the event by clearing their personal schedules. One of the more extreme anecdotes I encountered involved a moderator who

would "just change [their] sleep schedule" to better sync up with the event schedule. Another offhandedly mentioned that "since I have insomnia, I'm there for the European stuff no-one else can do." In any case, moderators coordinate to tell each other when they are available (or not) to monitor chat, with the intent to have some overlap to allow moderators to leave to take care of themselves, or simply take a break to focus on the games rather than watching chat alone. These events are huge time sinks, and over the course of the weekend a moderator may well spend upwards of ten hours per day on the event. Part of this is that moderators need to show up early to the event, because some viewers will show up before the stream starts, and viewers will stay after the stream ends, which means some moderators must keep an eye on chat after its official stop as well.

> It's more than a full time job during an event. I track my uptime, the hours I spent working on [a large recurring event] for two years, 2016 and '15 I think? I usually had roughly 140 hours of uptime in an event that ran for a week, for seven days...I [needed to] sleep, and I had university, I was up for all but like twelve hours. I slept for four hours a day at best. It was stressful. It's why I always suggest having more than one head moderator, because a single person can simply not take the load. Especially for a 24/7 event, it's just impossible. And if they don't take the load, or if there's not enough people, then they burn out and the quality of the management goes down.

3.2.2 During the Event

When the event proper starts, moderators load up different tools and applications. Common ones include Discord, opened to the event moderator group; a window with either the Twitch page itself, or one of several moderator-focused Twitch chat clients that enable multiple chats to be placed side-by-side in the same window; control panels for bots like moobot; Twitter or Reddit to monitor other chatter about the event; and personal programs such as a music player or a simple text editor.[5] This enables moderators to respond quickly to queries, to monitor multiple simultaneous chats—useful if the same video has multiple

[5] See Figure 3-4 for some sample screenshots of a moderator's setup.

alternate streams, as is the case with some major tournaments—in one window, to change bot settings on the fly, and a text editor allows a moderator to catch new spam as it flies by.

What mods are looking for are patterns within the chat. These patterns generally come as memes or spam. Spam was not well-defined by the moderators I interviewed, but can be taken to mean any message intended for multiple users to copy and paste to as to quickly repeat it in a very short span of time, while memes are a broader category of short, repeatable message. Spam in esports chat often involves the use of global or FFZ emotes, juxtaposing both the surface level meaning of the emotes-as-images with what they mean to esports fans, with some accompanying message, whether done in regular text, ASCII, or Unicode. Twitch esports chat spam changes extremely rapidly. Some spam is a response to the action happening on stream; others might be akin to fan chants, proclaiming support for one team or another.

Moderators are not paying attention to what individuals are saying in chat. Rather, they are paying attention to the patterns and rhythms of chat: which messages get repeated, if they are relevant to the stream or if they are irrelevant spam, and most crucially, if they are offensive, disruptive, or otherwise exhibiting a moderation "red flag". Over time, experienced moderators are able to predict which spam is most likely to get picked up and repeated. "When you watch chat for so long you pretty much know what's going to happen when it happens. It's like the Twitch chat whisperer, you just know everything." This sense for Twitch chat allows more experienced moderators to warn newer moderators what to look out for, and for them to react quickly to head off spam that might encourage more and more-offensive messages from being posted.

The types of offensive spam that moderators encounter are ever-changing. The most easily-identified offensive spam feature slurs or other offensive terms. These terms might be spelled plainly, or by using letter-substitution via using similar-looking glyphs, Unicode or ASCII symbols. More involved offensive messages involved the use of ASCII or Unicode to create offensive images, or the use of emotes (default, subscriber-only, or added by extensions such as BTTV or FFZ) to create offensive phrases and imagery. Moderators also described a class of offensive spam designed to trick viewers unfamiliar with American cultural norms into repeating offensive phrases, showcasing the sometimes creatively devious

nature of this spam.

> Twitch chat has a thing where it tries to trick other people into being racist instead of the original commentor. It's basically trolling people into being racist...As with all groups of people if it becomes too large, it's just a mess of idiots, and Twitch chat, if it sees something, it will most likely copy-paste it within seconds...There's been times where I've banned people for saying horrible things, and the person actually didn't know what they were copy-pasting at all.

For popular spam written in languages other than English, a moderator may try to tap others with other language skills in order to ensure spam remains slur-free. In my observations, I saw calls for Russian-speaking moderators, as well as German, Czech and Hungarian; knowledge of multiple languages remains a valued skill for moderators, especially as this sometimes corresponds to wider timezone coverage. Otherwise, moderators use tools such as Google Translate to try and get a quick sense of what the message means. Some non-English spam is so common that moderators who do not otherwise speak those languages do not need to check its translation to know what it means, and know when to remove it.

A large chat may be set up with several modes aimed at fractionally slowing down chat and protecting it from the most easily exploitable methods used by spammers to dominate. This would be turning on slow mode and a chat delay for moderation. Slow mode means that each user can only post messages after a certain time delay, preventing a single user or group of users from quickly overwhelming chat. Chat delay means that all messages have a small time delay, with moderators seeing them before they are passed to the public stream. This small delay allows moderators, and bots with moderator status, to act on messages ideally before they are even seen.

The first barrier to disallowed chat messages is AutoMod, which catches messages in its filter and prevents them from hitting chat at all, instead passing them to moderators for manual approval. Notably, a message held for human review by AutoMod will not show up for users at a whole, meaning that even without anyone giving AutoMod explicit 'allow' or 'deny' feedback, there is some level of chat filtering that will occur. The second is for a moderator, on spotting a disruptive message, to manually add it to a bot blacklist. This first

short-term action is generally done in a user-friendly bot with a GUI, such as moobot. There are two reasons for this: a GUI may allow for faster response as it can be opened and readied in a separate window, some mod hierarchies are organized such that most mods will have edit permissions for these less-powerful but easier-to-operate bots, and halting the exact spam message quickly limits its spread.[6] The longer-term solution, if the blacklist does not halt the spam, is to work this message into more powerful filters using regular expressions. Badly-configured regex strings have the potential to filter too many messages, turning the chat against the moderators, or to filter too few, allowing prohibited messages through; such was the case with one observed incident, where a typo meant that a regex filter meant to catch messages using a certain combination of punctuation was, in fact, catching no messages at all.

If possible, other moderators then encourage chat to continue by engaging in non-offensive memes or playing games with chat. This may also be done during down-time, such as scheduled ad breaks, or technical issues. This can also be a planned and delegated role; if moderators know a minority figure will be on screen, they may set up specific filters (for example, if a female commentator is part of the broadcast team, they may prepare stricter filters targeting sexist language) and be prepared to deploy a moderator specifically to distract chat when they are on-screen. This distraction generally involves trying to steer chat in the direction of less spotlighting and more benign behaviour; one example given to me was trying to engage chat in the 'golden Kappa test' or to ask them to 'build a rainbow ladder'. Both of these are well-known Twitch chat games, which essentially demonstrates participants' knowledge of Twitch by trying to trigger a known easter egg or by using special ASCII characters, respectively. The intent is to divert attention from the screen and to give chat something to do other than point out the minority status of the person on screen. Simultaneously, the users that originated this now-blocked spam may be timed out or banned, depending on severity.

Once the new filters are in place, moderators look out for signs that users (particularly those that originated the spam) are testing the boundaries of the new filters. One moderator, who said they were normally loath to "make an example" of a user, singled out these

[6]Few of these methods are backed up by data or evidence proving their efficacy, but through practice and observing the outcomes of adopting certain moderation actions over others, Twitch moderators have developed sophisticated techniques to regulate chat.

boundary-testing users as particularly dangerous "since they know how we blacklist things, they know how to edit it just right so it'll get through again. Those are the people I throw a day ban on. I'll @ them[7] after I ban them, 'Congratulations, you found a way to get past the filter, here's your prize.' That way chat knows they shouldn't do it." Moderators are very harsh towards these kinds of probing attempts, since they signal a conscious and deliberate attempt to test the abilities or attention of the moderators. If the moderators are sure that the popularity of that particular piece of spam has waned, it may be removed from all blacklists to stop false positives. Again, it is important to keep in mind that it is not necessarily the content of the spammed message that warrants its removal, but the fact that it is being repeated so much that no other messages can get through.

In a particularly dire scenario, moderators may limit the rate at which messages can be posted, using a more restrictive slow mode, r9k mode, a stricter AutoMod setting, follow-only mode, or sub-only mode. r9k mode disallows repeated identical messages, which forces spammers to change their messages before posting and thus slowing the flow of spam. Stricter AutoMod settings catch more and more messages, although an overly-strict AutoMod setting can actually become detrimental since it is not nearly as configurable as either ohbot or moobot; higher settings both catch more messages and broaden the net to catch more types of content. Follow-only mode restricts chat to those who have followed the channel for some amount of time, which slows the flow of newcomers jumping in and also stops banned users or other bad-faith actors from creating new, throwaway accounts to spam the channel. Sub-only mode only allows subscribers of the channel, those who pay money to support the stream, to chat. This is seen as equivalent to turning off chat altogether, since so few users subscribe to event channels. Because chat is the primary source of interaction between users and an event, and is vital to the cultivation of an organization's brand and fanbase, turning it off is a last resort.

3.2.3 Cleanup

Once an event ends, some moderators might stay around until the majority of the users have actually left, since it is common for people to hang around after the stream ends. Aside

[7] On Twitch, typing `@username` pings a user with that username, highlighting the message for them.

from that, there is little to do after the event. "The head mod will always be like, 'Good event guys, thanks a lot, our next event is going to be at so-and-so.'...At the end of the event, we go through all the commands in moobot and disable them so the people don't call in the chat, they don't prompt the command and get information from a month ago." However, the bulk of post-event work comes in the form of dealing with unban requests. Moderators may get private messages on Twitch from users that have been banned; since users are not always told who has banned them, the chances of a banned user contacting the moderator who was responsible for their ban are extremely slim. With a tool like Logviewer, any moderator can check a user's history and see what messages they posted up to their ban, and see if moderators have left additional notes about that user. Without access to Logviewer—for example, when I participated in moderation, a recent Twitch update had broken Logviewer functionality for the new Rooms feature—moderators must rely on each other's memories of the event to decide whether or not to unban a user. With luck, a moderator might have had the foresight to keep logs, generated from some other chat client, of the event; post-event logs are useless as deleted messages are not preserved even for moderators.

Moderators may also experience an uptick of angry messages left by irate users. Learning to handle these among moderators is a little-discussed part of becoming a moderator, as the practice of sharing these angry and sometimes threatening messages is the most common form of emotional and mental care in moderator circles. Some users are persistent: "we laugh about the death threats. There was, somebody posted death threats to almost every single ESL mod at an event two or three weeks ago and we were laughing about it in mod chat." Though moderators know they are not meant to share private messages, in these circumstances it is very common and forms the basis of moderator-specific memes and humor. Counter-trolling or baiting these angry users is considered an acceptable response. "Sometimes it can even be funny because they're so aggressive for things that are so negligibly unimportant. Often what I did was I had my fun with them, I pretended to be super nice."

Compensation may be given to the moderators. This is so rare as to almost be unheard of, and even then almost never comes in the form of money. If compensation is given, moderators may be given tickets for future events with VIP access, invitations to exclusive

afterparties, or "swag bags" of branded merchandise. Some moderators, who self-identify as fans of the esport for which they work, view the opportunity to meet esports celebrities as a form of compensation in itself. However, even these non-monetary forms of compensation are still incredibly rare.

3.3 The reality of event moderation

The section above assumes an ideal scenario, but not every event goes as planned. To begin with, not every event that is streamed on Twitch has a human moderation team, let alone a moderator liaison. With the development of AutoMod and the wide availability of automate moderation bots, such as Moobot, an event's organizers may decide to simply set up one or more of these tools as the extent of their moderation setup. While I am unsure of how common this is as a moderation strategy, it happens enough that one moderator I spoke to said that they had previously tried to advise event organizers on their moderation setup, but stopped because they were worried their advice would be taken as an insult.

Rarely, human moderators might be called upon to provide damage control for an event while it is ongoing. I was an observer in one such instance, where a representative from an event asked for help moderating their channel, and some volunteers responded. The speed and surety with which the moderators started trying to implement their tools, bringing in bots and manually taking action in chat, suggested this was not an unheard-of situation. The moderator in question had even set their username to be bright yellow, in order to stand out better in chat, so that whoever controlled the broadcasting account could spot them more easily in the chat and therefore, grant them moderator status more easily.

Even with a human moderation team, present at the very beginning of the event, there is no guarantee of training or adequate coverage of the event. Furthermore, not every organization will provide guidelines for their moderators, and even those with codes of conduct or similar documents for their chats may not have them prominently displayed, once again shifting the burden of both enforcement and education onto the moderators. The size of a team may also cause issues. Somewhat counter-intuitively, though, my interviewees mentioned that an overly-large team could cause more problems than a small one, since a small

team could still moderate effectively given knowledge and familiarity with automated tools. Suggestions of good numbers for an event moderation team ranged from "three or four" to "twenty", but those who advocated larger moderation teams also pointed out that this would make the moderation team "as big as the broadcast team", and therefore be somewhat impractical.

The issue they identified as hitting large moderation teams the hardest was communication. This was vividly displayed during DreamHack 2016, where a black professional Hearthstone player, TerrenceM, was playing, and Twitch chat devolved into a constant stream of racist commentary. Not only were the mods at the time unable to control the situation, some allegedly joined in and offered to unban those users who had been participating in the racist messages (Filewich, 2016). Less dramatically, though, large moderation teams that are a result of organizers indiscriminately adding all applicants may open the door for deliberate sabotage, as different groups of moderators conflict with one another. One moderator recounted such a situation:

> We have this...really big community of streamers. They have a monopoly on everything but they don't have a single decent moderator, but they still want to have the power...because they have access to the [biggest event] channels themselves on Twitch, because they stream from them, they decide, "Oh look, it's these guys again, they're moderating the chat but we want to have a monopoly. Let's demod them all." Obviously we message Twitch...[Twitch] has to mod us again, tell them not to do it again, they say, "Oh sorry, it's not us, it was just some new guy who wanted to do this, he's just handling the camera and got near the computer and decided to have fun." And this happens at every event.

This is one example where having a large group of moderators does not lead to better moderation. Such large groups may be a sign of indiscriminate hiring, heedless of existing social dynamics between groups of moderators with different styles and philosophies. A larger group of moderators, lacking clear guidance from the event organizers or a single respected head who can organize and enforce guidelines within the moderation team itself, can be a recipe for disaster. This is part of why moderators create channels for communi-

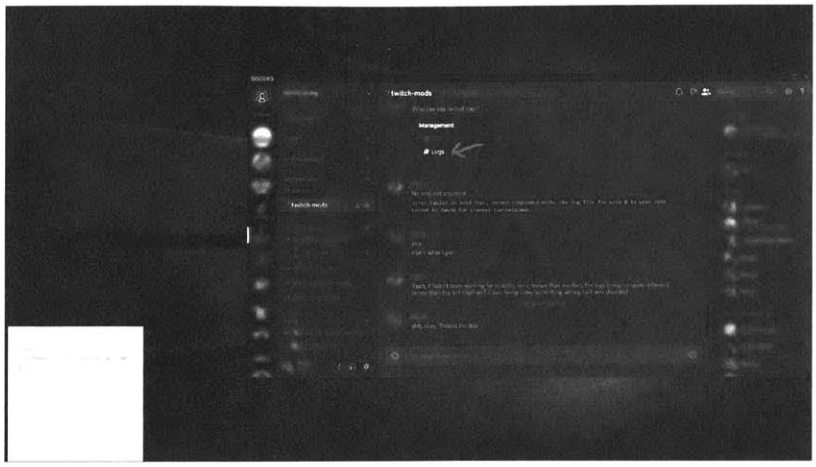
cation amongst themselves, and one more experienced mod said "the thing I take pride in having changed is the culture of communication within the moderation communities. Back in the day there was very little to no communication at all...I tried to enforce always having the mods online while they were moderating, that was actually one of the core rules we moderators had. If you were moderating, please be in the group and be contactable and read what I write in that group, if I messaged you please respond, because there was no way to keep a moderation team cool if people were just uncontactable." These attempts to create clear lines of communication do not always extend beyond the moderation group, however. Many of them said that contact with event representatives was rare, and if they could contact a person, there was no guarantee that they would be responsive, or have access to critical controls such as access to the broadcasting account. Some moderators also pointed out that they could help with the event beyond dealing with Twitch chat.

> A good example of that is [an event], two years ago. There was a kid up in the front row with a sign that said, "They're going A" and on the back, "They're going B". So he would actually be flipping the sign and trying to cheat for his team. That was one of the few times we actually got a hold of production almost immediately. Security ran down and got the kid, took away his sign, almost threw him out. And that's one where—*we need this all the time.* There was one other thing, it was too long ago and I can't remember, some UK event. There was some drunk dude that we could see when they panned the camera over to the crowd that was just off his rocker, half-naked with a horse mask on. We were like, "Uh, we can see this dude here we can see on stream, maybe call security?"

The emphasized section indicates that clear communication to people on the ground, working at the event, is rare for moderation teams. A few moderators said that often, their point of contact was the event's social media manager. This was because the social media manager, in charge of making live updates on the event's accounts and creating the corresponding graphics and short video clips, was also often watching the stream and staying constantly updated. However, because different events have different organizational structures, that position or work may be done by one person, a few people, or no-one in particular,

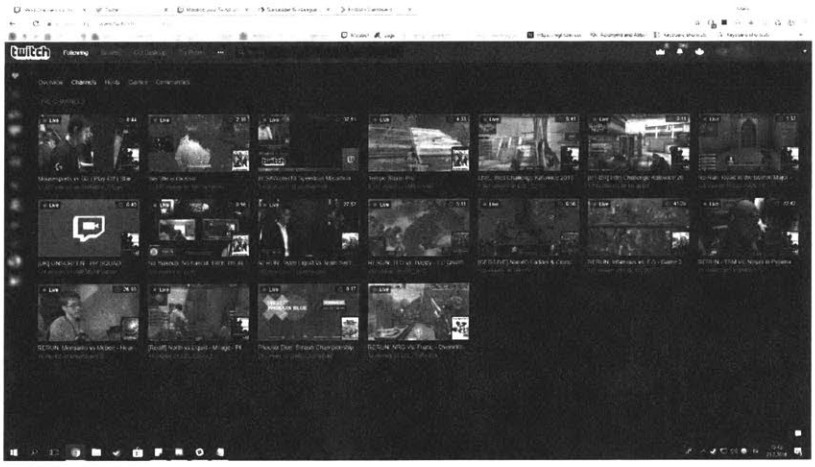
further complicating communication between moderators and event organizers.

The reality of event moderation is not a clear hierarchy, with volunteer moderators below event organizers. Rather, lines of communication are fraught. Volunteer moderators are at once given considerable leeway with regards to judgement calls, yet little to no power to communicate to other event staff. At once visible and invisible in chat, event moderators fit uncomfortably into the hierarchies of event organizations even as they are clearly apart from general spectators.



(a) Leftmost monitor.

(b) Middle monitor.



(c) Rightmost monitor.

Figure 3-4: An example of a moderator's triple-monitor setup.

Chapter 4

The Social World of Moderation

4.1 The life of a moderator

How does a user make the jump to becoming a moderator? The majority of my interviewees indicated that they started as moderators for a particular streamer's channel, rather than starting with event moderation. Through a combination of repeated and frequent visits to a single stream, and a similarly sustained history of answering questions or providing helpful criticism to the streamer or their moderators, became noticed and singled out as particularly helpful regulars. One moderator got their start after pointing out problems in a stream's bot setup; another acquired a reputation as a helpful regular for CS:GO streams by answering questions about event schedules, scores, and other basic information. This method, of picking known community members, is consistent across all of the interviewed moderators. At TwitchCon 2016, the moderator team for CohhCarnage, a popular streamer, discussed the qualities they looked for when recruiting a new moderator.

> "[It's] how we choose our mods, it's somebody who's spent a lot of time in the channel, a lot of time talking in the channel, and being, in our case, someone with good vibes and a good countenance that is safe-for-work. Just following our rules, being a member, and supporting the team...Mod selection is a biased process. Don't think that it's not. If people don't know who you are, and don't know what you're about, then you're probably not going to make it to mod

status."

The other, less-commonly discussed necessary condition for becoming a moderator is the presence of ample free time. The three situations my interviewees highlighted were unemployment, being a student, or being injured, thus freeing them up to invest an unusually high amount of time in watching and participating in Twitch chat.

The jump from community moderator to event moderator is quite small. Some of the interviewed moderators had previously been involved in setting up or running grassroots esports tournaments, and therefore their involvement in Twitch event moderation was an offshoot of their work for the tournament. Others were community moderators for esports-adjacent streams, such as moderating for esports professionals, who were then invited to work for a particular event. This movement was not all one-way, either, with some event moderators taking on additional work as a community moderator after exposure to a team or individual esports players after working esports events. Nevertheless, informal channels of recruitment, mostly consisting of calls for moderators on private moderator-only groups set up on platforms outside of Twitch or individual recommendations, seems to be common and standard practice for event moderator recruitment.

Some larger, more established esports event moderation teams might have a more recruitment process, involving written applications alongside informal recommendations. However, one head moderator admitted that the application was less important that simply observing newly-recruited moderators at work, saying that "on those applications they can write whatever they want, but if they [don't] do any moderation well then I won't have them on my team for longer than a week." The use of applications can also create a lot of extra work for head moderators, especially for popular events that might attract many eager applicants.

From a technical point of view, becoming a moderator is very simple. An existing moderator or the broadcasting account can use a simple chat command, `/mod username` to make someone a moderator; removing moderator status is similarly easy, `/unmod username`. Moderators gain a small white-on-green sword badge next to their username to indicate their status in chat. For this reason, moderator status is sometimes referred to as a "sword" or "badge".

A user with moderator status then has a different experience of Twitch chat. The most

apparent is that two buttons appear to the left of every user's name in chat. These allow moderators to ban or unban, and timeout for ten minutes, that user. Additionally, moderators can set chat to different modes. These modes include slow mode, which forces users to wait for a given amount of time between posting new messages; followers-only mode, which limits chat to users that have followed the stream for a given amount of time; r9k-beta mode, which only permits unique messages (and therefore cuts down on repeated or spammed messages); and in extreme cases, subscriber-only mode, which only allows subscribed followers to post to chat. This last one is considered a last-resort measure for event channels, since there is little to no reason for a user to subscribe and pay a monthly fee for an event channel, and consequently turning on sub-only mode for an event channel is like hitting the off switch for chat.

What kind of training is provided to a new moderator? The prevailing method essentially throws a newly-minted moderator into the deep end, relying on their natural trepidation and hesitance to radically overhaul the existing system to avoid major issues. A few moderators remembered a more hands-on mentorship, which affected their approach towards new moderators. One said that "One of [the stream's] more senior mods at the time messaged me and set me up, gave me some advice and rules and stuff...And I try to do the same thing with people who become mod, give them the same thing he gave me." This trepidation, or on the flipside, the confidence moderators feel for taking the initiative on moderation action, versus following others' leads, forms part of how moderators describe their position within moderation hierarchies. Some of the moderators that I interviewed considered themselves senior in some channels or for certain esports events, and junior in others, precisely because in more unfamiliar channels they did not feel comfortable taking a leading role in those moderation teams.

Because moderators, as discussed, tend to already be familiar with the moderation team, they are aware (or are quickly informed) of the existing team's hierarchy, generally from head moderators telling them who to turn to in case of an emergency. The moderators at the top of this hierarchy, whether officially titled head moderators or those wielding authority granted through reputation and seniority, act as mentors to new moderators. New moderators are expected to lurk, watch, and listen to more experienced moderators, and to follow in their

footsteps, though these expectations are rarely explictly stated. One interviewed moderator remembered consciously modelling their judgement calls and actions on what other moderators were doing, saying, "When I first got modded, I only banned the obvious, like people who would drop the n-bomb. That was about all I did for a little while. Then I started looking at things, like, 'I would time that out.' And about three, four seconds later, I'd see [the head moderator] time that out. I did that for a couple weeks, then I was like, 'I've got it,' everything I wanted to do he was doing."

Aside from being given moderator status, new moderators are granted access to moderator-only discussion groups. These fast-paced chat rooms function as a combination training ground, philosophical discussion space, and support group for moderators. While none of the interviewed moderators spoke of any formal training program or systems, they did say that these moderator-only spaces were where discussions about moderator policy happened. "We pretty much talk about everything. It's very open, and it's not just the head moderators or the chiefs or whatever their title is. We all discuss, all of the moderators, about policies, ethics, how we should do things, why we do them, how we could be more effective," recounted one moderator. This, combined with more experienced moderators checking on newer moderators' actions during events, forms the basic mentoring model for a event moderator on Twitch. By being in these discussion groups, new moderators learn by watching what more senior moderators consider good or bad actions, and get feedback as to which of their actions were appropriate or not. In my observation, even experienced moderators would frequently check in with the other moderators to ask for second opinions, such as whether or not a user should be banned for having a slur in their username.

As newer moderators gain experience and confidence, they may take on more and more responsibilities and seek out more knowledge related to moderation. The proliferation of scripts, bots, and other automated tools means that basic programming knowledge is important for moderators to learn; in particular, knowledge of regular expressions is greatly valued. Nearly all the moderators I spoke to talked about the importance of learning regex, or at least understanding what it was and how it functioned. Moderators may also specialize, "[incorporating] what they are studying, for example, into moderation... A lot of the up-and-coming moderators last year have been young kids doing what they like, things they've been

doing in school." Others read up on fields they believe may have use to them, "whatever you can imagine, from economics to history to psychology." This extra knowledge or experience can be used to stake out a niche for themselves and their own moderation styles. This is one way in which a newer moderator can make a name for themselves, and the desire to create better tools for their work can be a powerful motivating factor for a moderator to teach themselves how to program.

Finally, moderators who end up heading whole teams of moderators must take on managerial duties, managing teams of people and liaising between an event organizer and their team, taking over training and mentorship duties, in addition to performing moderation work. Seniority and authority gained from increased responsibility also means that, in the more informal world of the Twitch moderation community, one's word gains weight in disputes between moderators. In the words of one moderator, "I was one of the most, if not the most, experienced person on the mod team. So if I just stated the result of an argument, that was it, the discussion was usually over." Part of their authority comes not only from their history of work, but from the fact that moderators with a long career are able to remember and bring up justifications and anecdotal evidence for setting moderation policy, further lending credence to their arbitration. Senior, more experienced moderators' work also extends beyond the technical and managerial, and into decisions involving moderation policy, philosophy, ethical concerns, and pitching in with recruitment or training. Some larger, established moderation teams might have training documents, pointing new moderators towards certain tools and settings for them, as well as giving examples and best practices for them to follow. These extend beyond decisions about content moderation. For example, one extremely common recommended practice was turning on two-factor authentication, a matter of online security. Again, all of these best practices documents are assembled from moderators' own knowledge and circulated within a relatively small closed group.

Given all of this, what might make a moderator leave? In a direct mirror to one of the conditions for becoming a moderator, a lack of free time might lead to a slow decline in the amount of work a moderator can do, which in turn may herald their departure from the moderating community. Overwork, too, was cited as a cause for moderators leaving. Recalling one high-profile departure, one moderator said that "he quit moderation...a year

ago, because he had had enough, because he pretty much overworked himself extensively for free. And, you know, you can only do that for so long until your body caves in and you just give up." This was a rare acknowledgement of the lack of compensation for taxing moderation work directly leading to a well-regarded moderator leaving. Even more striking is its description of burnout, a spectre that hung over discussions about overwork and moderator mental health in my interviews. While none of the interviewed moderators admitted to feeling burnt out or even significantly affected by their work, all of them would admit that they had heard of burnout or overwork contributing to someone else's decision to leave.

Another is a disconnect between themselves and the communities to which they feel they belong. This makes sense, given that many of the moderators I interviewed said they identified as a fan of the esports scene before they became a moderator, and that the community (whether of moderators or of that particular esport) was a reason for them to persist as a moderator. One powerful anecdote came from a longtime *Counter-Strike: Global Offensive* moderator, who recalled an incident that nearly made them stop moderating.

> There were some kids on screen. There were eight or nine-year old kids on screen...Chat said some fucking heinous shit. I'm sorry for my language, but that's tame compared to the things they said. I actually typed in chat, "I'm done, I can't do this any more." Because chat just repulsed me...I slept 39 hours in the past nine days and I was just too tired for it and I saw people saying, "I would rape that kid," or "Pedobear come here." Just the worst shit....that was the first time I reached the level of disgust. It made the bile in my stomach churn.

The emotional and mental drain on moderators does not only come from viewing and having to then remove the worst of Twitch chat. Every single moderator I interviewed had said that they received harassment, abuse, and death threats; all of them told me this with a laugh, signalling that it was so common an occurrence as to become routine and in some cases, the basis for jokes. There are few other sources of support for moderators aside from the aforementioned moderator-only chat groups, with some saying that they felt they could not discuss their work because "it's really hard to give a shit when, you know, it's the Internet...[people think] it doesn't matter what people say." In other words, the general view

that talk is cheap online is harmful for moderators trying to have their work taken seriously. Even within moderator circles, all of them talked about the need to develop a thick skin or to otherwise get used to constant abuse as either a matter of course for all moderators, or as a requirement to persist in this role.

Moderators might also leave because of conflicts with other moderators, collectively referred to as 'drama', and sufficiently dire disagreements may force even formerly well-respected moderators to quit their teams. 'Drama' forcing moderators off of their teams was something that hit some interviewees hard, especially if they were head moderators. One head moderator, who could not work one event due to schedule conflicts, blamed himself for not being there when a subsequent disagreement in his team lead to someone leaving. "There was just that leading figure missing who could explain how and why the rules were as they were, and as a matter of fact one person then left the moderation team. That was slightly unfortunate, that was just because I didn't put enough effort into keeping the moderation team cool and controlled." Instead of pointing at the harsh work conditions, lack of support or compensation, or other structural factors that stress moderators, he attributed the loss of this moderator to his failure to lead the team.[1]

Leaving the world of volunteer moderation for paid full-time work is also a fraught path. It is currently vanishingly rare for a volunteer moderator to transition in to a paid moderation job; at the 2017 TwitchCon panel, "Keeping the Peace through Moderation", the panelists could name "about five" moderators who managed to get a full-time paid position that was directly relevant to Twitch moderation. The more common path, according to my interviewees, was that volunteer moderation served as a way to network with people from the esports scene, with the hopes that these connections would result in a job working for various event organizers. The resulting jobs were, anecdotally, most commonly in production or social media management. While this makes some sense—Twitch moderation requires one to be technically proficient in working with the Twitch platform, and with essentially managing one's own public reputation as well as upholding the event's desired public brand

[1] This is actually something of a recurrent theme: moderators seem well aware of the structural conditions arrayed against them yet take personal responsibility for much of what goes wrong e.g. burnout, quitting, disputes, or chat going nuclear. Unsure how to link it, but I feel this is a topic for elaboration in the 'reality ensures' chapter.

or values—neither makes direct use of the expertise these moderators develop.

4.2 The relationship between moderators and Twitch chat

In order for moderation to function at all, a working relationship of trust must exist between the moderators and the chats for which they moderate. The relationship between a moderator and the chat which they oversee is not reduceable to a one-sided relationship of power, or even a single stable mood. What is undeniable is that, for these large-scale event chats on Twitch, there is constant friction and tension between these two groups, while the ghost of other involved parties—event organizers, the personalities at the head of the most popular Twitch streams, and Twitch staff—also influence the ongoing and dynamic relationship between mods and the chat. It is critical to remember that moderators are drawn from Twitch, are embedded within esports communities, and often are moderating events for which they are a fan. They understand the culture of Twitch, and, as a consequence of their work, have had close experience with how it has evolved over the years, as well as a deep understanding of how best to engage with chat. In other words, they make full use of sociability, "the collective purpose of a community, the goals and roles of the individuals in a community, and policies generated to shape social interaction" (Preece, 2000) to conduct their work.

In section 3.2.2, I previously outlined what red flags moderators look out for, and the steps they take to address such messages. Understanding how these moderators view and interact with Twitch chat is key to understanding the underlying logic behind these actions. The interviewed moderators had a variety of metaphors with which they described chat, the most vivid of which was describing it as "like a caged beast." Recurring themes in all these descriptions centered around its knee-jerk contrarian behaviour, extremely short attention span, surprising cunning, and the contrast between individual chat users, as expressed through private messages to individual moderators, versus chat as a whole. This last point is important: while moderators might variously describe individuals as "kids", "trolls", or more neutrally "users", they seem more concerned with the mob entity made up by the mass of users, which is what I have referred to as 'the chat'. This understanding of chat shapes the moderators' actions as well as moderation policy more generally.

According to the moderators I talked to, the prevailing wisdom for managing Twitch chat was to be relatively permissive and to allow chat to have its fun, spamming memes and reacting, provided they did not start using offensive language, or having one piece of spam predominate and take over chat. Continuing on from the "caged beast" metaphor, one moderator explained chat's behaviour by saying that "if you cage it in too small, it will get really angry and try to find any attack vector, in order to be as offensive and hateful as possible. But if you set the bounds quite loosely, they will tone down their tone as well." Often, the trade-off that moderators decide upon is to allow disruptive spam, knowing that removing it would provoke an outsized reaction. One moderator volunteered an anecdote explaining this mindset:

> An example I can give is in Rocket League chat, which is the official world championship channel by Psyonix. People have just been spamming 'gg'.[2] And they just spam it all day long and it's annoying as hell, completely ruins chat. But if you were to ban that, the backlash you would get would be way above and beyond what you would get as a chat-ruining experience right now. So you just leave it and let them spam their 'gg' and are happy that it isn't anything racist.

These moderators understand their role and work as guiding chat more than controlling chat. One said that "chat has the attention span of a goldfish. You don't have to go heavy-handed and blacklist every little thing and knock out every little spam, because in five minutes they'll have something else they want to spam." Rather than trying to set strict filters that catch every instance of impermissible speech, moderators rely on their knowledge of the rhythms of Twitch chat in order to knock certain spammed messages out of circulation, or to suppress it temporarily. Because of the way in which repetition and spam drives the pace and patterns of chat, moderators are not just concerned about offensive chat messages (although they do, of course, remove them). Rather, they are particularly focused on offensive chat messages that have a memetic, viral, or otherwise transmittable quality to them.

[2]Meaning "good game", sometimes used to politely thank an opponent for a match, but in this context meant sarcastically to deride someone for an easy win.

When asked what the point of Twitch moderation was, almost all of my interviewees said that they understood the point of moderation as keeping chat fun and ensuring viewers enjoyed the event. The one moderator who did not say that this was the primary purpose of moderation instead saw their job as part of ensuring streamers and channels kept to Twitch's Terms of Service, thus protecting them from possible removal from the platform, but they still expressed a moderation philosophy strikingly similar to the others. Granted, part of this may stem from the fact that the world of large-scale esports event moderation is quite small; certainly, there were a few key members of this community that my interviewees greatly respected and pointed to as the source of their own values, policies and philosophies regarding Twitch moderation. We can see that moderators want to promote good experiences for viewers, partially as a goal in and of itself, but also because they are aware that this makes chat far easier to regulate than a stricter, more impermissive style.

The other piece of this puzzle comes from the fact that Twitch chat is very aware that they are being moderated, and are also aware of some of the tools and methods with which moderators regulate chat. For example, Moobot is not designed solely for moderation, and it is easy to set up for any channel; therefore, Twitch chat users are quite familiar with its operation since any of them could set up the bot themselves to look at its default filters and settings. This knowledge is not always expressed maliciously. A moderator pointed out that knowledge of bot-based moderation is itself the basis for jokes for chat, saying,

> They actually @ moobot, they're like, 'moobot you shouldn't be blacklisting that', you can see them throw fake insults at it. They've become more aware of the tools we use and how we set those tools. If we set slow mode on for more than ten seconds they'll be like, 'oh slow mode!'

The anecdote also shows that chat, collectively, can be very astute when it comes to figuring out how the moderators of the channel are operating. Again, as previously mentioned in section 3.2.2, some bad-faith actors know enough about how bot filters work to try and modify their messages to get around the filter. By harnessing the observations of many viewers at once, chat can figure out different settings and use this as a new point of focus. This same astuteness when it comes to figuring out moderator actions is used by the moderators

in order to change chat norms. Without having to explicitly state rules for chat, or by referring users to guidelines or similar, moderators can influence chat norms simply through enforcement. One moderator described the long process of teaching some sections of event chat to get used to new chat guidelines.

> The first time he [hypothetical user] gets timeouted, maybe he won't understand why. And he tries this again. By using a lot of different software, we see he's doing this a second time and we timeout him again for one day, or half a day. And you can see that people evolve, on the small scale—we have a chat with 300,000 viewers, obviously we can't check everyone. But I started checking, ten random people, what they do, what they did, did they adapt. And you can clearly see the pattern. They adapt...in 90% of the situations, they know you can't say this any more.

What this moderator banked on was the ability of chat to first notice this new enforcement, and then to realize that this new boundary was being regularly enforced. Equally, the moderators rely on the bevy of third-party tools at their disposal to see the extent to which these shifts are taking hold.

Moderators must be aware of other considerations when it comes to anticipating flare-ups or fights in chat. I have previously mentioned that moderators may choose to take extra precautions when they know that minorities will be shown on screen, by adding extra filters, pre-emptively knocking out certain spam patterns, and being ready to redirect chat towards less harmful games. However, for these events, moderators also have to keep apace with scandals, fights, or other points of disagreement within the scene specifically. More generally, moderators also have to be attuned to wider Twitch cultural shifts. During IEM Katowice's finals, one flagged situation involved insults and provocation between subscribers to two popular Twitch streamers, DrDisRespect and Forsen. They were easily identified because they were using custom emotes that they acquired by becoming paying subscribers to their respective streams, and using them to mock subscribers or followers of the other. The moderators indicated that a fight between DrDisRespect and Forsen's followers was quite common, since both had such huge subscription bases. They also suggested that the lax

moderation found on popular Twitch channels was a major contributor to general problems of Twitch chat toxicity.

Just as these moderators understood and saw 'the chat' as an entity distinct from the individuals that comprise it, Twitch chat thinks of moderators as 'the mods', a collective entity that obscures individual moderator. Jokes or popular spam messages in Twitch chat about that tension between a channel's moderators, and a sufficiently-cunning chat, abound. Some are so common they have been collected on different sites that help users spam these messages even more easily. Along with these messages and their endless variants, was the frequent use of the phrase 'Nazi mod', to refer to what chat users believed was overly heavy-handed moderation.

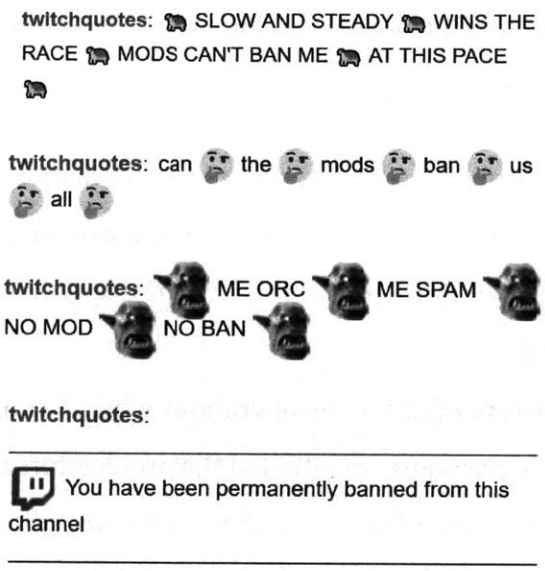


Figure 4-1: Some examples of popular mod-related spam.

As we can see in figure 4-1, three are about testing or poking fun at the limited attention span and efficiency of the moderators, while the last aims to impersonate a moderator as a joke. During my experience moderating during IEM Katowice, many users were not aware of how to tell a mod from a regular user, with some claiming to be a moderator on the basis of having a Twitch Prime badge[3], or being able to send coloured messages. Others, on seeing messages deleted in real time, would attribute these actions to other non-moderator users

[3]This badge is available for anyone who pays for Amazon Prime.

in the channel. More commonly, though, they would openly discuss whether they thought a human moderator was present in the room or if it was merely an automated filter; the consensus seemed to be that it was a bot. This confusion, with human moderators' actions mistaken for bot actions, was frequent enough that in the moderator-only Discord server, some mods would jokingly introduce themselves as 'often confused for a robot'.

However, breaking their silence to participate in chat is not always a good choice for a moderator. Even if they engage with chat, their status means that they sometimes voluntarily bar themselves from participation. Moderators are wary of being drawn into public debates or conflicts, even if they want to intervene, with one saying that "sometimes I feel like I want to say something but I can't because I'm a mod." Indeed, when I was moderating, a few users started to argue about why there were so few women in professional gaming; when I said that it was unlikely that genetic factors were to blame for this disparity, one user started to complain about "mod baiting", in other words, saying that the moderators were intentionally stirring up conflict. Instead of understanding my opinion as something that I held as an individual, which chat seemed to approve of when it was strictly in the realm of discussing games, my dissenting opinion was understood as automatically done in bad faith or meant to entrap them.

In other words, making moderators visible in chat is a careful act that is constrained by informal rules and norms. It is something moderators learn to navigate based on previous experience and knowledge of Twitch chat's general relationship to moderators, and to that specific chat's disposition towards this particular moderation team. Outside of making their individual presence known, these moderators were also careful to present the team as a unified force. Discussing inter-mod conflict publicly was considered extremely damaging to the moderation team as a whole. One well-known and well-watched event, Games Done Quick, has its volunteer moderators sign a non-disclosure agreement in order to prevent leaks; while an extreme example, it shows the reticence of moderators and organizers to make the details of their decision-making public.

For the interviewed moderators, the ideal Twitch event chat was one where permissible memes and spam proliferated, without a single spam or type of spam dominating, and where users either did not notice chat moderation or did and paid no attention to it. They saw the

free creation, remixing, and repetition of memes as a key part of what made viewing esports on Twitch so enjoyable, and so it was important to them to create an environment where new memes could be made.

The worst case scenario was a chat in open revolt against moderation; that is, instead of chat complimenting the video stream and reacting to it, chat is arranged against the moderators, possibly making a game out of testing the limits of the moderators (e.g. boasting about being kicked multiple times, repeating memes related to moderator abuse or "Nazi mods"). Successful chat moderation on Twitch requires the tacit approval of a majority or, at least, silent plurality of Twitch chat users; when users feel dissatisfied with the moderators, the situation can easily escalate, requiring more heavy-handed tactics to control the rate of messages coming through.

4.3 What is good moderation?

Being modded adds a little green sword-shaped badge next to one's username in chat. This serves to mark out mods and has become slang: to "get your sword" is to become a moderator. There is social cachet to having that badge by one's name, since it is also the closest thing to an authority figure in Twitch chat; in community spaces this also acts as a mark of recognition, since moderators there are so often promoted from regular, generally well-liked members. Several moderators described this badge as something that was generally coveted; the general interpretation, joking or otherwise, of mods as power-hungry or otherwise able to throw their weight around no doubt adds to the perceived value of moderator status.

My interviewees identified two primary types of 'bad moderator': vanity moderators, or those who got moderator status with no intention of doing moderation work, and rogue moderators, which are compromised or malicious moderators. They also highlighted 'badge hunters', which I class as a subset of vanity moderator; these are users who seek moderator status on as many channels as possible while not putting in a commensurate amount of work. As of the time of writing, Twitch provides no other easy way for streamers to distinguish users in chat other than granting them moderator status, and therefore a moderator's sword-

badge. Therefore, streamers occasionally "sword" their friends, significant donors, or other users who they feel deserve recognition, but are not expected to perform any of the work of a moderator. However, this phenomenon is rare in event moderation since there is little to no reason why an esports event organizer would need to distinguish chat users in this way.

4.3.1 Rogue moderators, online security, and handling threats

Security issues are one of many considerations for event moderation teams. If any moderator's account is not sufficiently secured, it might be compromised, and since being a moderator is currently an all-or-nothing proposition, a compromised moderator account is capable of significant harm. These rogue accounts can mass-ban or -unban users, destroy bot settings, cause that moderator to be removed from the channel, and generally cause chaos. Therefore, the adoption of security measures such as two-factor authentication has become a de-facto standard for the large-scale event moderators I talked to. A few recalled being the target of attempts to compromise their accounts. One moderator remembered an incident where someone "leaked [their Twitch account] plain-text password in a crowded chat I was modding. I had no idea who they were, all they did was post my plain-text password, and I had to frantically change every single one of my passwords. I was lucky, he got it wrong by one character. But that same night, he did that to other mods that I knew. He would, for example, get on their account and start banning everyone and get them unmodded." Another described a very sophisticated hacking attempt, done to advertise a website.

> One of our moderators got hacked...He was specifically targeted, they wanted to make ads [originating] from a moderator...Moderators can't demod other moderators, only a streamer can, so we couldn't do anything. And we tried to timeout his followers who pasted the same advertisements. But we couldn't, because they had a script that purged all of our timeouts from his account. So it was really pretty creative, I don't know if they reached their goal, but it was thoughtful. It was planned ahead. But we reacted pretty fast, we contacted Twitch and after three minutes they banned that account. And afterwards we banned all the people who followed him. All of this took two or three minutes.

Responses to rogue moderation accounts are still relatively rudimentary. Tools such as moderation logs or mass demodding scripts can help as an emergency response; one moderator who recalled trying to deal with this problem before the existence of these tools said that, "for a while there was no moderation logs on Twitch, meaning we didn't know who was banning whom, so if I saw a lot of bad timeouts or bans the only thing I could do was write in the group, "Hey everyone, I'm seeing a lot of bad timeouts, can we maybe change that, who has been doing that, please stop that." If they wouldn't read the group there was nothing I could do. If there were hacked accounts...the only thing we could do was unmod everyone, and see when it stops. And that person was the person that got hacked." Prevention, in the form of taking online security measures, is the preferred method of dealing with the threat of compromised accounts.

However, the moderators did not mention being taught to take online security measures, whether relating to keeping their account secure or keeping their personal information private. When I directly asked one moderator, he replied, "No, never actually talked to any mods about that kind of thing. Never heard anyone bring it up either. The only reason I ever really thought of it maybe was because we laugh about the death threats." Even then, adoption of security measures, even in the face of constant threats, is not always uniform. The same moderator said, of using safety or privacy measures, "I hope other people do. Personally I don't care."

It was not lack of knowledge that kept him, and other moderators, from taking these measures—when asked what should be done to ensure one's online safety, he rattled off a short list of things to do: "Oh, you know...never put your real name on stuff, never have your email account tied to something they could look you put for, never have your email address public." It was a cultivated sense of stoicism or apathy; as previously mentioned in section 4.1, all the moderators I talked to were quick to say they were personally unaffected by the many threats they received. In a notable point of comparison, they constantly justified their reactions by pointing to the belief that others had it worse than they had. For example, when asked about the abuse they had received, one moderator replied, "Oh yeah. So many death threats!...Doesn't bother me whatsoever...I've never really had anyone say, 'I'm going to find you, I'm going to stalk you,' heavy stuff. I'm sure there are people out there who have

been doxxed or something like that, and do have a fear of that, a fear for safety, but I'm a little unique in that sense." Yet another moderator specifically brought up an instance where someone attempted to threaten them by doxxing, to which they reacted with that same kind of learned apathy.

> I know there was one guy who that was trying to get my information, and he went to this old website...that used to have leaked passwords and stuff. He got my password off that site once and tried to track me and say, 'I know your passwords and emails,' and stuff like that...I actually just found it funny because it didn't really matter any more, and he said stuff like, 'I have your address'...I don't really try to hide that either, it's all public. I feel like I'm in a really safe country so I don't have to worry about anything happening, like getting swatted. I don't think any moderators have ever been swatted...he got it from a site that posts leaked stuff that you have to pay for, so he paid money to get my information and then tried to track me with it, and I just found it funny that he paid money for that. Because he tried to pretend that he was a hacker, but I knew exactly where he was getting the information from.

It should be noted that, despite the stoicism demonstrated by the moderators I interviewed, the moderation community frequently engaged in caring mental health and emotional support behaviour for one another. This included a channel specifically for posting cute animal pictures, and another for sharing memes or to joking about stressful aspects of moderation. I also saw frequent occurrences of moderators checking in on one another, talking each other through stressful situations or checklists of best practices for dealing with harassment on Twitch. On this last subject, they were extremely knowledgeable: they detailed specific actions that the user in question would have to do to build an actionable case against their harassers for Twitch to become involved, as well as suggesting other steps to protect themself while also reassuring the user in question that it was fine to step back from moderation to engage in self-care. Some of the moderation guideline documents I saw also included links for hopefully de-stressing content, such as a 'cute animals' link, in their 'Resources' sections. Clearly, in private, moderators felt more comfortable expressing frustra-

tion, fear, and other negative emotions towards the work that they had to perform. In short, despite the affected air of nonchalance that moderators display when asked about the stresses of moderation, especially that which stems from the constant environment of abuse, they are well aware of the toll it takes on those volunteers who perform this work.

4.3.2 Badge-hunters and proper moderation values

As for vanity moderators, the moderators I interviewed seemed to reserve special ire for badge-hunters, or more generally, those users who seek moderator status with no intention of putting in the work. Also called badge collectors, they were described as people who "basically just try to become a moderator on every chat they can get hold of, but they don't do anything." This problem was not isolated to new moderators. Another moderator pointed out that badge-hunters were sometimes already established and well-known in the moderation community, saying that "we have a lot of moderators that are already known, but they try to get as many swords as they can to just collect them for some reason, and you can see they try to participate in the most events. But they don't do anything. It seems to me like they don't care about their reputation." Indeed, some praised the creation of moderation log tools, which keep a record of every action taken by every moderator, as a way of seeing which moderators were pulling their weight.

At the same time, a few of my interviewees were very critical of the perceived motivations behind users who volunteered for moderator positions; these individuals also tended to be the ones who admitted to initially having the "wrong reason", in their own words, for becoming a moderator. These wrong reasons included believing that moderation was primarily about "the power to remove anything from the chat that is not welcome", or for the more selfish reasons of "[getting] that extra popularity in chat and [having] your messages more noticed." They were more united on their descriptions of good moderators, and good moderation motivations: wanting to help out the channel, streamer, or event, advancing the interests of the broadcaster, and ensuring that chat-goers enjoyed themselves without that enjoyment coming at the expense of either the event organizers or other chat users.

Interestingly, some of my interviewees maintained that this proper attitude towards motivation could not be taught. Simultaneously, others described a transition where "you

kind of learn to love the community as a mod instead of as a community member", and clearly, those moderators who admitted to seeking moderator status for selfish reasons have changed their minds over the course of their career. While they were quick to talk about changing minds on issues of policy, they did not raise as many examples of changing minds on the issues of norms and values. This suggests to me that the proper or good moderation values are being taught to new moderators, but through informal channels. One clear way is the mentorship method by which these moderators learned to carry out their work: by consciously modelling themselves on role models, it would not be surprising if they also picked up their role models' values. Others may have changed their mind after witnessing moderator disagreements or discussions about particular moderation policies.

Indeed, some moderators talked about changing minds on the nuts and bolts of moderation policy regarding permanent bans, both as a new moderator and as a mentor. However, as they elaborated on the reasons behind why they thought the way they did about permanent bans versus temporary timeouts, their reasons started to diverge: some believed it was too heavy handed, some cited personal experience that it simply created more work dealing with unban requests in the long run, others talked about mentors who explained their own values and beliefs to them. For example, one moderator remembered being taught not to hand out many permanent bans, saying, "I got taught when I joined moderation that we don't ban people...I am not afraid of banning, but I think that's alright not to do, timeout is enough. People tend to learn someday and if you ban them, they just create another account and you have to ban them again, there's no point." In short, they clearly spoke about a change in values that occurred alongside a change in policy, justified by personal and passed-down precedents guiding anticipated outcomes and their estimated potential for future backlash.

This was most apparent over the contentious issue of emote bans. Twitch has long had a reputation, and a problem, with toxicity around the use of global emotes depicting minorities. Some of the more commonly used emotes, for these offensive purposes, are TriHard and cmonBruh, both depicting a black man; HotPokket, depicting a woman with dyed blue hair; and Anele, a man in a turban. These global emotes are taken from well-known Twitch streamers or Twitch staff, but have been appropriated and used for their surface representation of these particular minority groups in order to express offensive sentiments, in

ways that are harder for filters to handle. The issue of offensive emote spam, particularly spamming TriHard, came up during the 2016 DreamHack Hearthstone tournament, where a black professional Hearthstone player was on-screen and Twitch chat responded with a torrent of racial abuse, including spamming TriHard. More recently, professional Overwatch player Félix "xQc" Lengyel was released by his team after an incident where he spammed "TriHard 7" when reporter Malik Forte was on-screen.

Figure 4-2: Global Twitch face emotes often used in offensive messages. From left to right: TriHard, cmonBruh, HotPokket, Anele

Multiple moderators told me that the norm a few years ago was to blanket-ban all use of specific emotes that were used for offensive purposes. They were also very emphatic in telling me that this was the wrong choice. From my observations of moderator-only spaces, this is still an ongoing debate, but many of the more respected moderators now err on the side of not blanket-banning emotes. In the words of one moderator,

> There were some moderators who thought that auto-timeouting TriHard, just TriHard without any context, was a good idea. And it was a 'good idea' for a pretty long time…Now we have ohbot that checks the context. But the general rule should be that any Twitch global emote should be allowed if you're not using it in a bad way. It's always about the context, it always should be.

The various rationales given for context-checking and allowing global emote usage, even if it made cutting down on racist or offensive spam harder, included the argument that global emotes should by default be permitted, and that it was not a solution to remove emotes depicting minorities from usage, given that the majority of face emotes are of white men. However, this decision to check context, which can be automated because of the availability of regular expression filters, is also in keeping with these moderators' general belief that permanent bans are last-resort tools; in short, instead of suppressing usage of these emotes, they wanted to discourage their use in specific instances. In the words of one moderator, moderation was "more about giving [them] the chance to say nice things instead of bad

things." The decision made by these moderators to use more complex context-checking filters, which screen what is being said before and after the emote itself, should be understood as a policy decision made from repeated observation of the effect of blanket-ban policies, in keeping with existing moderation philosophies, and implemented in automated tools as a result of the affordances of said tools.

Because few esports organizations or game developers give their moderators chat guidelines, codes of conduct, or other guiding documents, moderators have had to develop standards for chat on their own. Even as they acknowledge that they are often the sole arbiters of acceptable behavior in chat, many of the expressed discomfort with their position. When asked if these organizations should be leaving these judgements to moderators, one replied quite forcefully, "No! I would not trust the majority of people." Others spoke of their attempts to remain "impartial", and of the responsibility they felt that they had to model good behavior. Another directly told me, "You also have to understand, for me it's really important, that Twitch chat moderators are not—they can be wrong...I think you should look at this too in your studies because it's important how people make mistakes. In terms of how they moderate or why they do this. Some people are power hungry, some people want to show off how they do bad things." Moderators also are aware of the ethical dimensions and implications of their work. In addition to the reticence I described regarding their positions as rule-makers, one moderator I talked to also expressed wariness over the ways in which third-party tools, especially automation, could be deployed.

> Well it's become easier [moderating], but the problem also is, where do we cross the line with how easy we can make it?...It's also about how far can you take it. There's a fine line between, how much information should we keep about these people in the chat? If a guy is being an absolute asshole in one chat, should we ban him from another chat? ...[A hypothetical banlist] would get abused, no doubt about it, immediately I would guess. Somebody doesn't like your opinion? Well I have the power to ban you from thirty chats. You're not allowed to be here any more. That's why we actually talk a lot of ethics nowadays and ways to do things better. It's so easy to go too far as well, I feel. A lot of moderators are like, they want to have full control, they want to moderate everything, but I

feel like that's not the correct way to go.

Ethical considerations also factor into moderators' interpretation and handling of new platform policies. That is, when Twitch announces new guidelines for streamers, moderators are quick to adapt, since they know they will now be responsible for enforcing them even if they may not be directly monitored. This is especially interesting in light of the fact that not all moderators agree with all of Twitch's Community Guidelines. In March 2018, Twitch clarified their community guidelines by saying that "as a streamer, you are responsible for the content on your stream." As the community of moderators was quick to note, this could easily mean that streamers could be held responsible for the actions of their chat if they automatically displayed or responded to messages from said chat, for example if they had an automated system that allowed donors to display any message of their choice or if their stream included a live view of Twitch chat. A discussion broke out, with some moderators arguing that streamers should not be held responsible for the actions of their viewers, while others argued that, due to their influence and reputation they held, streamers should be held accountable for how they shaped their chat. What is notable about this is the fact that this discussion centered less on what moderators would have to do, but rather on how the relationship between streamers, their moderators, and their chat should be configured. The moderators of this community were able to comment and theorize, as well, on the historical developments of Twitch chat that might have led to the current state of Twitch chat culture, in order to justify or strengthen their arguments.

Chapter 5

Moderation Futures

In the scramble to comply with pressure on platforms to be seen regulating, who is left to actually carry out enforcement? As our desire for greater and more visible moderation rises, there is no guarantee that our understanding of the work of moderation will grow alongside it. We began by noting that the pace of scandals around platform moderation seems to be picking up. How might the future of moderation develop, at least for this group of moderators? And what can we learn from this case study, especially with regards to what we as online users, citizens, and participants ought to be asking?

I began this project with the intention of making clear the complexity of moderation work, and in particular its cultural, communicative and social aspects. Yet at the same time as I acknowledge that there are issues of scale at play for the largest social media platforms that means it is hard to directly translate the volunteer model to moderation for them. However, I still believe that understanding the work done by volunteer moderators for their communities, both communities of users and communities of moderators, is important in developing future support for them as well as considering what design for community growth might entail.

5.1 Twitch, esports, and event moderation

As the esports scene changes, so will Twitch and its livestreaming competitors; the desire to draw new audiences and sponsors is pulling at the scene and expectations of what is appro-

priate within it. Yet considerations of moderation rarely appear in popular discourse surrounding it. Moderation only tangentially enters the discussion when it is centered around the racist, toxic behaviour of both esports pros and of Twitch chat, and generally framed in the context of how the organizations involved—game developers and esports teams—choose to handle the situation. Platform exclusivity deals, such as ESL's exclusive streaming deal with Facebook for some of their largest DOTA 2 and CS:GO tournaments, further complicate the future development of the esports moderation landscape. There is no guarantee that these platforms have the kinds of moderation tools and affordances that allow volunteer moderators to carry out their work. Indeed, in the case of Facebook, their real-name policy and policies against allowing multiple accounts means that moderators cannot rely on anonymity to provide some measure of safety against reprisals.

At the same time, competitors to Twitch are pushing more transparent codes of conduct, most notably Microsoft's Mixer. Yet even when platforms champion transparent or easy-to-understand policy, I do not see any of them champion moderation features, unless it is to highlight automated chat message removal tools. Given all that moderators do, and the specific tools that allow them to do better work that have nothing to do with automated message removal, the fact that this seems to be the only aspect of moderation that platforms are willing to work on displays a deep gap between volunteer moderators' practical experiences and what developers are ready to give them as tools.

Yet, it is important to keep in mind the potential power of a moderating community. Specifically, the communication networks and relationships formed between these volunteer mods has led to the formation of an organized force of workers, despite their lack of compensation. i believe that the value of moderation work, as chronically undervalued as it is, is nonetheless recognized as important to the formation of lasting communities, no matter how dimly. What is concerning is the ongoing lack of attention paid to this group of expert practitioners. Why do we see so little done to innovate for moderators, except from other moderators?

5.2 Transparency, accountability, and reporting

Calls for transparency and accountability for platforms around their moderating decisions are becoming increasingly common, as scandals about moderation, or the lack thereof, keep cropping up. However, demanding increased transparency is not enough. All else remaining equal, when we demand transparency for invisible work performed by invisible workers, we wind up seeing nothing at all. Who do we want to be more transparent? About what? And for whom? It is insufficient to demand knowing what is being removed and why. Instead, we need to start asking about the *philosophies of moderation* that a given platform holds. All platforms moderate, even, and especially, those that insist that they are neutral (Gillespie, 2018). While it is important that we know what is being removed, calls for transparency need to encompass more.

As my interviewees have shown, continued enforcement is more than merely that. Our focus on what gets removed exposes an appetite for understanding moderation only as it pertains to the most visible aspects. Equally, it means we will always remain mired in a kind of naive fascination with individual points when what we need to pay attention to is the trajectory of moderation. In other words, we need to pay attention to the ways in which policy is enforced, and how this enforcement creates and contributes to cultural and social changes on the platform itself. We also need to understand that communities hosted on platforms respond differently and nimbly to policy changes enacted by platforms. Finally, we need to understand who and what operates the mechanisms of moderation in any given space, and the conditions under which they labor. That is to say, we must expand our question to include what that moderation is doing, and how it is being done.

I would argue that we, as users, deserve to see more than just notifications of content removal. We need to have a moderation trajectory from the platforms on which so much of our online social interactions occur. A moderation trajectory would include a philosophy or some kind of articulation of what that platform believes to be good or proper moderation, how it ought to be achieved, and a roadmap of broader goals, both policy-based and concrete objectives, that can be achieved. A trajectory of moderation would be a statement of moderation's purpose with respect to the socio-technical context of the platform.

There are a few important implications of demanding moderation trajectories as a part of the push for transparent platform moderation. Firstly, and most obviously, in order to tell us what their moderation trajectory is, platform operators have to also know what it is. This requires forethought. It means they must approach moderation proactively rather than retroactively, to have had a plan in place before public outcry or evidence of missteps, wrongdoing, or scandal.[1]

Secondly, a trajectory—a projected future course—requires more than platitudes. Having a trajectory allows us to understand the moderation decisions of a given platform operator in the context of past precedent and future goals., It means giving users the ability to make sense of what actions and policies were present, are currently implemented, and how they may change in the future. Again, it emphasizes moderation as a proactive series of decisions with social repercussions beyond the immediate consequences of removal. This is important because it will also require that we, the public, be able to fit both ongoing enforcement and points of failure into this trajectory. To criticize failures of moderation when we are unsure of what we want, and what is being offered, as a moderation trajectory is one thing; it will be another to criticize a sub-par trajectory of moderation.

The trajectory would make visible the work; next, we must make visible the human realities of moderation. The workers are a fundamental and inseparable part of online moderation. With respect to volunteer moderation, labour conditions are sometimes reduced to the presence or absence of compensation, but it does not tell the whole story. Volunteer moderators can be a well-networked and well-organized labour force. The absence of formal support for volunteer moderators should not be confused with there being *no* support; the lack of visible organization does not mean there is *no* organization.

If a given platform is going to offload any of the burdens of moderation onto its users, then we must see that what they allow in terms of self-governance and community organization and moderation, as an integral part of its moderation trajectory. If they expect

[1] Demanding a transparent moderation trajectory means thinking of moderation as *going somewhere*. I purposefully choose this term over a more common descriptor, such as a moderation 'strategy', in large part because I do not want to emphasize militarily-minded metaphor: there is not always an 'us' against a 'them', and in any case the makeup of these groups are constantly in flux. I believe 'trajectory' is a better term because it more strongly emphasizes the fact that there is a path to be laid, which has been neglected in favour of thinking about reactive action. Granted, it implies a single or a clear path; nevertheless I think it is the better term.

self-governance, what has been provided, both in the design of the platform, and in the expressed values and norms that the platform expects these moderators to uphold? The two are not mutually exclusive, and in order to understand and critique moderation we must be conversant in both.

In an ideal world, we would have proper support for moderators, professional or otherwise. Support would not just be limited to compensation: it would include the recognition of moderators as an important constituency of a given space, and both the desire and ability to take their voices seriously and to provide the conditions, material and otherwise, to allow them to do their work to the best of their abilities. Platforms would be invested in publicizing cogent, coherent plans of moderation, with track-records to bear them out, and be able to elaborate on the philosophies guiding said moderation actions. As users, we would be aware of and be able to participate in the work of moderation: not just the labour of flagging, reporting, and so on, but the ability to decide on governance structures and access to the information required to make informed decisions.

Realistically, though, I cannot expect even some plurality of users to bring themselves to care so strongly about online moderation. People go on Twitch to watch livestreaming, not to watch other people watching. But this is precisely why it is so important to recognize volunteer moderators as a distinct group of invested users, occupied with organizing, technical, communicative, and social work. So long as moderation remains invisible, the actions of a few will continue to have an outsized impact on our online lives. I am not trying to say that volunteer moderators are either a force for good or a userbase to be feared. Rather, they are motivated people who have found ways to take some small measure of control back over their online lives, and of the communities to which they are tied. The human costs of the work are exacerbated by the situation in which they, and we, are mired; not just costs to the workers, but the costs of missteps, failures, and the subsequent need to be on guard against malicious action. The structural factors arrayed against them are undoubtedly vast, but not necessarily insurmountable. The work is hard, but worthwhile. And the first step is to acknowledge what has already been done.

Bibliography

Braun, Joshua. 2013. Going Over the Top: Online Television Distribution as Sociotechnical System: Online Television Distribution. *Communication, Culture & Critique* 6:432–458. URL `https://academic.oup.com/ccc/article/6/3/432-458/4054515`.

Filewich, Carling. 2016. "Enough is enough": Confessions of a Twitch chat moderator. https://www.gosugamers.net/hearthstone/features/39013-enough-is-enough-confessions-of-a-twitch-chat-moderator.

Gasser, Urs, and Wolfgang Schulz. 2015. Governance of Online Intermediaries: Observations from a Series of National Case Studies. *SSRN Electronic Journal* URL `http://www.ssrn.com/abstract=2566364`.

Geiger, R Stuart, and David Ribes. 2010. The Work of Sustaining Order in Wikipedia: The Banning of a Vandal. *CSCW* 6:10.

Gillespie, Tarleton. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* New Haven & London: Yale University Press, 1st edition.

Grimmelmann, James. 2015. The Virtues of Moderation. SSRN Scholarly Paper ID 2588493, Social Science Research Network, Rochester, NY.

Humphreys, Sal. 2013. Predicting, securing and shaping the future: Mechanisms of governance in online social environments. *International Journal of Media & Cultural Politics* 9:247–258.

Jenkins, Henry. 2008. *Convergence Culture.* New York University Press, revised edition.

Jeong, Sarah. 2015. *The Internet of Garbage.* Forbes.

Kerr, Aphra, and John D. Kelleher. 2015. The Recruitment of Passion and Community in the Service of Capital: Community Managers in the Digital Games Industry. *Critical Studies in Media Communication* 32:177–192. URL `http://www.tandfonline.com/doi/full/10.1080/15295036.2015.1045005`.

Kerr, Aphra, Stefano De Paoli, and Max Keatinge. 2011. Human and Non-human Aspects of Governance and Regulation of MMOGs. 23.

Klonick, Kate. 2017. The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review* URL https://papers.ssrn.com/abstract=2937985.

Kou, Yubo, and Xinning Gui. 2017. The Rise and Fall of Moral Labor in an Online Game Community. 223–226. ACM Press. URL http://dl.acm.org/citation.cfm?doid=3022198.3026312.

Kraut, Robert E., Paul Resnick, and Sara Kiesler. 2011. *Building successful online communities: Evidence-based social design.* MIT Press.

Latour, Bruno. 1992. 'Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts'. In *Shaping Technology/Building Society: Studies in Sociotechnical Change*, ed. Wiebe E. Bijker and John Law, 225–258. Cambridge, MA: MIT Press.

Massanari, Adrienne. 2015. #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society* 19:329–346. URL http://journals.sagepub.com/doi/10.1177/1461444815608807.

Matias, J Nathan. 2016. The Cost of Solidarity: A Quasi Experiment on The Effect of Joining A Strike on Community Participation, in the 2015 reddit Blackout.

Niederer, Sabine, and José van Dijck. 2010. Wisdom of the crowd or technicity of content? Wikipedia as a sociotechnical system. *New Media & Society* 12:1368–1387. URL http://journals.sagepub.com/doi/10.1177/1461444810365297.

Postigo, Hector. 2016. The socio-technical architecture of digital labor: Converting play into YouTube money. *new media & society* 18:332–349.

Preece, Jenny. 2000. *Online Communities - Designing Usability, Supporting Sociability.* John Wiley & Sons, Ltd.

Roberts, Sarah T. 2012. Behind the Screen: Commercial Content Moderation (CCM).

Shaw, Aaron, and Benjamin M. Hill. 2014. Laboratories of Oligarchy? How the Iron Law Extends to Peer Production. *Journal of Communication* 64:215–238.

Silva, Leiser, Lakshmi Goel, and Elham Mousavidin. 2009. Exploring the dynamics of blog communities: The case of MetaFilter. *Information Systems Journal* 19:55–81.

Terranova, Tiziana. 2000. Free labor: Producing culture for the digital economy. *Social text* 18:33–58.

Thomas, Bronwen, and Julia Round. 2016. Moderating readers and reading online. *Language and Literature* 25:239–253. URL http://journals.sagepub.com/doi/10.1177/0963947016652785.

Twitch. 2018. Community Guidelines FAQ Update. URL https://blog.twitch.tv/community-guidelines-faq-update-a322c82b8038.